FACULTAD DE ESTUDIOS ESTADÍSTICOS

MÁSTER EN MINERÍA DE DATOS E INTELIGENCIA DE NEGOCIOS

Curso 2020/2021

Trabajo de Fin de Máster

TÍTULO: "Análisis de datos para segmentar a los clientes y aumentar la eficiencia de las campañas de Marketing"

***Alumno*: Maria Emilia Tocco**

***Tutor*: Ramón Alberto Carrasco y Rocío González Martínez**

Septiembre de 2021

UNIVERSIDAD COMPLUTENSE
MADRID

Índice

1. Introducción 1
   1.1. Contexto 1
   1.2. Justificación del proyecto 2
2. Objetivos 3
3. Metodologías y Métodos 4
   3.1. Metodología para el proceso de minería de datos 4
   3.2. Segmentación de clientes 6
      3.2.1. Algoritmos de Clustering 6
      3.2.2. Algoritmo K-Medias 9
      3.2.3. Modelo RFM 11
   3.3. Algoritmo a priori 14
4. Desarrollo del trabajo y principales resultados 16
   4.1. Comprensión del Negocio 16
   4.2. Comprensión de los Datos 17
   4.3. Preparación de los Datos 20
      4.3.1. Creación de nuevas variables 20
      4.3.2. Depuración de datos 21
         4.3.2.1. Tratamiento de datos atípicos variables de intervalo: 22
         4.3.2.2. Tratamiento de datos faltantes: 23
      4.3.3. Análisis exploratorio de los datos ya depurados 24
      4.3.4. Normalización de variables 25
   4.4. Modelado y Evaluación 26
      4.4.1. Modelo RFM 26
         4.4.1.1. RFM score en KNIME: 26
         4.4.1.2. Segmentación de clientes en base a las tres dimensiones del modelo RFM en R: 28
      4.4.2. Evaluación modelo RFM 33
      4.4.3. Segmentación de clientes en base a su ingreso, antigüedad y cantidad consumida en R. 41
      4.4.4. Evaluación segmentación según ingreso, antigüedad y cantidad consumida 45
      4.4.5. Primeras Conclusiones de las Segmentaciones 47
      4.4.6. Algoritmo a priori 49
         4.4.6.1. Mayores consumidores de vino 54
         4.4.6.2. Mayores consumidores de carne 56
         4.4.6.3. Mayores consumidores de fruta 57

4.4.6.4. Mayores consumidores de pescado 58
4.4.6.5. Mayores consumidores de dulces 59
4.4.6.6. Mayores consumidores de productos de bazar 59
4.4.6.7. Clientes con mayor valor monetario 60
4.4.7. Evaluación y conclusiones Algoritmo A priori 61
5. Conclusiones 62
6. Bibliografía 64
7. Anexo 67
7.1. Capturas configuración de los nodos de la depuración de datos SAS Miner. 67
7.2. Capturas configuración nodos utilizados en KNIME 68
7.3. Código R 76
7.3.1. Clustering 76
7.3.2. Algoritmo a priori 79

Índice de Ilustraciones

Ilustración 1 - Diagrama de flujo de la Metodología CRISP-DM 4
Ilustración 2 - Funcionamiento algoritmo k-media (Han, Kamber, & Pei, 2012) 10
Ilustración 3 - Ley de Pareto (el 20% de los clientes de una empresa generan el 80% de los ingresos) (Córdoba, 2011) 11
Ilustración 4 - Ejemplo RFM Score (Córdoba, 2011) 13
Ilustración 5 - Funcionamiento algoritmo a priori 15
Ilustración 6 - Gráficos de caja de cantidad consumida de cada uno de los productos de la base de datos 19
Ilustración 7 - Gráficos de barra de las variables relacionadas con el número de compras realizadas por los clientes 19
Ilustración 8 - Flujo creación nuevas variables en KNIME 20
Ilustración 9 - Gráficos de barras variables Age e Income 21
Ilustración 10 - Matriz de correlación entre variables 24
Ilustración 11 - Gráfica de barras, cantidad de clientes cada uno de los puntajes globales (RFM score) 27
Ilustración 12 - Gráfico de caja de las dimensiones del modelo RFM 28
Ilustración 13 - Gráfico de dispersión RFM (PCA) 29
Ilustración 14 - Gráfico de dispersión RFM (PCA) con visualización de 4 clústeres realizados con k-media 29
Ilustración 15 - Gráfica Elbow RFM en R 30
Ilustración 16 - Gráfica Average Silhouette RFM en R 30
Ilustración 17 - Gráfico de frecuencia de la cantidad de clústeres en base a la cantidad de veces que fueron elegidos 30

Ilustración 18 - Visualización de los clústeres RFM para k=4 33
Ilustración 19 - Gráfica de la silueta de los clústeres formados (RFM k=4) 34
Ilustración 20 - Gráfico de cajas de las variables Monetary, Frequency y Recency para cada clúster 34
Ilustración 21 - Visualización de los clústeres RFM para k=7 36
Ilustración 22 - Gráfica de la silueta de los clústeres formados (RFM k=7) 36
Ilustración 23 - Gráfico de cajas de las variables Monetary, Frequency y Recency para cada clúster (k=7) 37
Ilustración 24 - Gráficos de tartas de los clústeres formados y el RFM score 40
Ilustración 25 - Gráfico de cajas de las variables Ingreso, Antigüedad y Valor Monetario 41
Ilustración 26 - Gráfico de dispersión Ingreso, Antigüedad y Valor Monetario 41
Ilustración 27 - Gráfico de dispersión de Ingreso, Antigüedad y Valor Monetario con visualización de 4 clústeres realizados con k-media 42
Ilustración 28 - Gráfica Elbow para clústeres con variables Ingreso, antigüedad y Valor Monetario 42
Ilustración 29 - Gráfica Silhouette para clústeres con variables Ingreso, antigüedad y Valor Monetario 42
Ilustración 30 - Gráfica de barras número óptimo de clústeres 43
Ilustración 31 - observación clústeres en base al ingreso, antigüedad y valor monetario para k=4 45
Ilustración 32 - Gráfica silueta media de clústeres en base al ingreso, antigüedad y valor monetario para k=4 45
Ilustración 33 - Gráfico de cajas de las variables Income, Antiquity y Monetary para cada clúster 46
Ilustración 34 - Gráficos de tartas de los clientes "estrella" y "potenciales" en función de la segmentación RFM 48
Ilustración 35 - Gráficos de tartas de los clientes "Atención" y "Perdidos" en función de la segmentación RFM 48
Ilustración 36 - Frecuencia de los 10 ítems más frecuentes 51
Ilustración 37 - Frecuencia relativa (soporte) de cada uno de los ítems pertenecientes a la base de datos 52
Ilustración 38 - Resumen itemsets frecuentes 53

Índice de Tablas

Tabla 1 - Resumen explicativo de los distintos métodos de clustering (Ansari, 2021) ... 7
Tabla 2 - Descripción de las posibles variables a utilizar 17
Tabla 3 - Estadísticos asesor avanzado SAS Miner. 18
Tabla 4 - Estadísticos de sumarización variables de intervalo (nodo DMDB, SAS Miner) 21

Tabla 5 - Estadísticos de sumarización variables de clase (nodo explorador de estadísticos, SAS Miner) 22
Tabla 6 - Tabla para observar asimetría de las variables 23
Tabla 7 - Cantidad de observaciones con datos atípicos en las 5 variables que se van a utilizar 23
Tabla 8 - Escalas definidas para cada una de las 3 dimensiones, Recencia, Frecuencia y Valor monetario 26
Tabla 9 - Cantidad de clientes cada uno de los puntajes globales (RFM score) 27
Tabla 10 - Centroide de los clústeres formados en R para k=4 31
Tabla 11 - Centroide de los clústeres formados en R para k=5 31
Tabla 12 - Centroide de los clústeres formados en R para k=6 32
Tabla 13 - Centroide de los clústeres formados en R para k=7 32
Tabla 14 - Tabla con el mínimo, máximo y media de cada dimensión del modelo RFM en k=4 34
Tabla 15 - Observaciones pertenecientes a cada clúster con k=4 35
Tabla 16 - Tabla con el mínimo, máximo y media de cada dimensión del modelo RFM en k=7 36
Tabla 17 - Observaciones pertenecientes a cada clúster con k=7 38
Tabla 18 - RFM score para cada uno de los 4 segmentos encontrados 39
Tabla 19 - Resumen clústeres en base al ingreso, antigüedad y valor monetario para k=4 43
Tabla 20 - Resumen clústeres en base al ingreso, antigüedad y valor monetario para k=5 44
Tabla 21 - Tabla con el mínimo, máximo y media de cada variable y del RFM score en cada clúster 46
Tabla 22 - Observaciones pertenecientes a cada clúster con k=4 47
Tabla 23 - Itemsets frecuentes y su soporte 53
Tabla 24 - Reglas de Asociación para mayores consumidores de vino 54
Tabla 25 - Reglas maximales para mayores consumidores de vino 54
Tabla 26 - Reglas de Asociación para mayores consumidores de vino eliminando reglas redundantes 55
Tabla 27 - Reglas de Asociación para mayores consumidores de vino, pero sin tener en cuenta su hábito de compra 55
Tabla 28 - Reglas de Asociación para mayores consumidores de carne 56
Tabla 29 - Reglas maximales para mayores consumidores de carne 56
Tabla 30 - Reglas de Asociación para mayores consumidores de carne, eliminando reglas redundantes 56
Tabla 31 - Reglas de Asociación para mayores consumidores de carne sin tener en cuenta los hábitos de compra 57
Tabla 32 - Reglas de Asociación para mayores consumidores de fruta 57
Tabla 33 - Reglas de Asociación para mayores consumidores de fruta eliminando las reglas redundantes 57

Tabla 34 - Reglas de Asociación para mayores consumidores de fruta sin tener en cuenta los hábitos de compra 58
Tabla 35 - Reglas de Asociación mayores consumidores de pescado 58
Tabla 36 - Reglas de Asociación mayores consumidores de pescado eliminando las reglas redundantes 58
Tabla 37 - Reglas de Asociación mayores consumidores de pescado sin tener en cuenta los hábitos de compra 58
Tabla 38 - Reglas de Asociación mayores consumidores de dulces 59
Tabla 39 - Reglas de Asociación mayores consumidores de dulces eliminando las reglas redundantes 59
Tabla 40 - Reglas de Asociación mayores consumidores de dulces sin tener en cuenta los hábitos de compra 59
Tabla 41 - Reglas de Asociación mayores consumidores de productos de bazar 60
Tabla 42 - Reglas de Asociación mayores consumidores de productos de bazar eliminando las reglas redundantes 60
Tabla 43 - Reglas de Asociación mayores consumidores de productos de bazar sin tener en cuenta los hábitos de Compra 60
Tabla 44 - Reglas de Asociación clientes con mayor valor monetario sin tener en cuenta los hábitos de Compra 61

Índice de Ecuaciones

Ecuación 1 - Estadístico de Hopkins 8
Ecuación 2 - Confianza itemset 14
Ecuación 3 - Lift itemset 15
Ecuación 4 - Normalización Z-score 25
Ecuación 5 - Normalización Min - Max 25
Ecuación 6 - RFM Score 26

1. Introducción

1.1. Contexto

¿Qué es el Marketing?

Las personas en todo el mundo tienden a tener una limitada percepción de lo que es el marketing. En caso de preguntarles su opinión acerca de la definición de marketing, muchos (o la mayoría) responderían que es publicidad, promoción o creatividad. Estas respuestas no serían erróneas, sino muy limitadas con carencias de los factores que hay por detrás de esa publicidad. La publicidad es una de las tantas ramas del marketing, pero en un mundo cada vez más digitalizado, detrás de toda acción de la empresa, existe la importancia de un buen análisis de datos. (Portillo, 2021)

Según Kotler, *"en términos sencillos, el marketing es el manejo de las relaciones redituables con el cliente. El objetivo del marketing consiste en crear valor para los clientes y obtener valor de ellos a cambio."* (Kotler & Armstrong, 2012)

El marketing fue evolucionando a lo largo de la vida. En un principio, cuando no existía competencia, la función del marketing en las empresas era limitada, estaba orientada a la producción y se dedicaba a aumentar la eficiencia, a producir más y mejor. Se intentaba reducir gastos, aumentando los beneficios y la eficiencia.

Cuando empieza a existir competencia, aparece la preocupación por la calidad del producto y es entonces cuando el marketing se comienza a orientar al producto. Las empresas intentaban competir intentando ofrecer un producto con una calidad superior. Pero si todas las empresas competitivas aumentan la calidad del producto para intentar destacarse, termina siendo imposible diferenciarse únicamente por tener un producto de calidad superior.

Es entonces cuando nace el marketing orientado a las ventas. Aquí es cuando las empresas comienzan a realizar propaganda, publicidad. Esta fase del marketing parte de la idea de que, si se persuade a los clientes con una promoción intensa o una técnica de venta agresiva, el mismo terminará comprando mi producto.

De todas maneras, llega un momento que por más que se presione a los clientes y se promocione mi producto, ya no se vende. Es necesario llegar a una cobertura amplia, pero sin saturar el mercado. Invertir el dinero dedicado a las campañas de marketing de manera más eficiente. Surge entonces el marketing orientado al cliente, en dónde la prioridad se enfoca en el consumidor final. Las empresas comienzan a realizarse preguntas como: ¿Quién es el que compra? ¿Cuáles son sus necesidades? ¿Por qué está comprando el producto? El marketing identifica, orienta y estimula las necesidades y los deseos de los clientes para aumentar la demanda. Entonces el marketing satisface las necesidades de los clientes, presentándoles lo que ni ellos sabían que necesitaban. Se busca establecer una relación con el cliente. ¿Por qué? Si el cliente está satisfecho, va a seguir comprando, si no lo está, no va a comprar más.

Por último, surge la orientación de responsabilidad social del marketing, que nace de la idea de que satisfacer los deseos actuales de los consumidores puede implicar que no actúe con los intereses a largo plazo de la sociedad. ¿A qué se refiere? *"El concepto de marketing social señala que la estrategia de marketing debería proporcionar valor a los*

clientes de forma que conserve o mejore el bienestar tanto del consumidor como de la sociedad" (Kotler & Armstrong, 2012).

Estos cinco conceptos de marketing siguen siendo válidos al día de hoy y las organizaciones podrían utilizar uno o varios de ellos para diseñar y poner en práctica sus estrategias de marketing.

En el presente TFM, nos centraremos en el marketing orientado al cliente, poniendo como prioridad al consumidor final.

1.2. Justificación del proyecto

Como fue mencionado anteriormente, en un mundo cada vez más digitalizado, la importancia del análisis de datos en las campañas de marketing se vuelve cada vez más relevante. La facilidad para la obtención de datos a la hora de realizar campañas lleva a una utilización de los recursos mucho más eficiente.

En este TFM se pretende demostrar cómo el estudio y análisis de los datos puede proporcionar un impulso significativo a la eficiencia de una campaña de marketing, aumentando las respuestas o reduciendo los gastos. Se cuenta con los datos de una empresa minorista de alimentos, en donde nos dan todo tipo de información acerca de las características tanto personales, como de compra de los clientes. En este sentido, se busca entender y conocer a los clientes e identificar sus necesidades para optimizar la utilización de los recursos y lograr un máximo aprovechamiento de ellos.

"Sabiendo de antemano cómo piensa el cliente, uno podrá mejorar e innovar su estrategia. Se marca una tendencia en cómo el cliente ha actuado en el pasado o actualmente y permite marcar el camino probable que tome el consumidor en los siguientes días o meses." (Portillo, 2021)

Hoy en día existen muchos métodos de minería de datos posibles a ser aplicados en una campaña de marketing. *"La minería de datos puede ser considerada un súper conjunto de muchos métodos diferentes para extraer insights de datos. Podría implicar métodos estadísticos tradicionales y Machine Learning. La minería de datos aplica métodos de muchas áreas diferentes para identificar patrones antes desconocidos de datos. Esto puede incluir algoritmos estadísticos, aprendizaje basado en máquina, analítica de texto, análisis de series de tiempo y otras áreas de la analítica. La minería de datos incluye también el estudio y la práctica del almacenaje y la manipulación de datos."* (SAS, 2021)

Si nos centramos en el aprendizaje basado en máquinas *(Machine Learning),* los dos métodos más ampliamente adoptados son aprendizaje supervisado y aprendizaje no supervisado.

Los algoritmos de aprendizaje supervisado son entrenados utilizando ejemplos etiquetados, como una entrada donde se conoce el resultado deseado. El objetivo de los modelos de aprendizaje supervisado es aprender y descubrir los patrones que pueden predecir correctamente el resultado. En el caso de aprendizaje supervisado, siempre hay un conjunto de datos históricos etiquetados con una variable objetivo. Todas las variables distintas del objetivo se denominan predictores / características (variables *input*, independientes). Según el objetivo, el modelo identifica un patrón, establece una relación entre las variables *input* y luego usa la receta derivada para

predecir objetivos desconocidos en un nuevo conjunto de datos independiente. Algunos de los algoritmos de aprendizaje no supervisado pueden ser *Regresión Lineal o Logística, Random Forest, Gradient Boosting, Support Vector Machine*, entre otros. (Das & Cakmak, 2018)

Por otra parte, el aprendizaje no supervisado es una técnica de aprendizaje automático, donde no es necesario supervisar el modelo. No utiliza variable objetivo. En su lugar, debe permitir que el modelo funcione por sí solo para descubrir información. Se ocupa principalmente de los datos no etiquetados. En este sentido, la respuesta deseada no es conocida, por lo que no se puede utilizar la información del error explícito para ayudar a mejorar el comportamiento del algoritmo. El objetivo es explorar los datos y encontrar alguna estructura en su interior. Por ejemplo, se utiliza para realizar grupos de clientes con atributos similares (para que después puedan ser tratados de manera semejante en campañas de marketing). Los distintos segmentos de clientes deben ser heterogéneos, deben tener entre sí diferencias significativas que generen reacciones diferentes ante diferencias en los productos y/o servicios. Algunos algoritmos de aprendizaje no supervisado pueden ser mapas con organización automática, k-means clustering y descomposición de valores singulares. (SAS, 2021)

En el presente estudio nos centraremos en el aprendizaje **no supervisado**, teniendo como objetivo la segmentación de clientes aplicando análisis RFM y algoritmos de clustering, siguiendo la metodología CRISP-DM. La principal razón por la cual el proyecto se basará en Machine Learning no supervisado es porque a lo largo del máster ya utilicé esta base de datos para realizar el proceso completo de Machine Learning supervisado. Se intentó responder a un problema de predicción, en donde la variable objetivo era *Response* que tomaba el valor 1 si el cliente responde a una oferta de un producto o servicio y 0 en caso contrario. Se llegó a la conclusión que el modelo que mejor predice la variable objetivo es un modelo Random Forest correctamente tuneado. Para no repetir el proceso realizado a lo largo del curso, conocer la base de datos de manera completa y aprender más, he decidido focalizar el TFM en algoritmos no supervisados.

El trabajo está organizado de la siguiente manera:

2- Se presentan los objetivos del trabajo.
3- Una vez presentados los objetivos, se presenta la metodología empleada. En este apartado se realiza un breve marco teórico acerca de la metodología CRISP-DM en donde se detallan las distintas fases de la misma. Asimismo, se encuentran las definiciones de los algoritmos que se van a utilizar para segmentar a los clientes.
4- En el apartado “desarrollo del trabajo y principales resultados” se realizan todas las fases de la metodología CRISP-DM para mi conjunto de datos y se aplican los algoritmos de aprendizaje no supervisado.
5- Para finalizar se detallan las conclusiones extraídas tras la realización del trabajo, analizando el cumplimiento de los objetivos.

2. Objetivos

El objetivo principal del proyecto es realizar distintas segmentaciones de clientes con sentido de negocio, logrando conocerlos e identificarlos para aplicar de forma más

efectiva las distintas campañas de marketing. Se probarán distintas maneras de segmentar a los clientes para intentar encontrar la segmentación óptima.

Este objetivo se logra a partir de los siguientes objetivos secundarios:

1- Descripción y depuración de los datos.
2- Calcular el valor del cliente en base al modelo RFM.
3- Realizar análisis RFM para conocer cuáles son los clientes más y menos rentables, logrando asimismo identificar clientes que están a punto de abandonar la empresa, cuando antes eran buenos clientes. Se realizará el modelo convencional RFM.
4- Encontrar cual es la mejor manera de segmentar a los clientes en base a su ingreso, cantidad consumida y antigüedad.
5- Identificar, con la utilización del algoritmo a priori cuál es el perfil del consumidor de cada uno de los productos de la base de datos.

3. Metodologías y Métodos

3.1. Metodología para el proceso de minería de datos

Son tres las metodologías más utilizadas en minería de datos: KDD, SEMMA y CRISP-DM.

En el presente trabajo se utilizará la metodología **CRISP-DM.** La misma fue creada por el grupo de empresas SPSS, NCR y Daimer Chrysler en el año 2000. En la actualidad se considera la metodología más utilizada en el desarrollo de proyectos de minería de datos. CRISP-DM realiza el ciclo de vida completo de un proyecto en 6 fases: Comprensión del negocio, Comprensión de los datos, Preparación de los datos, Modelado, Evaluación e Implantación.

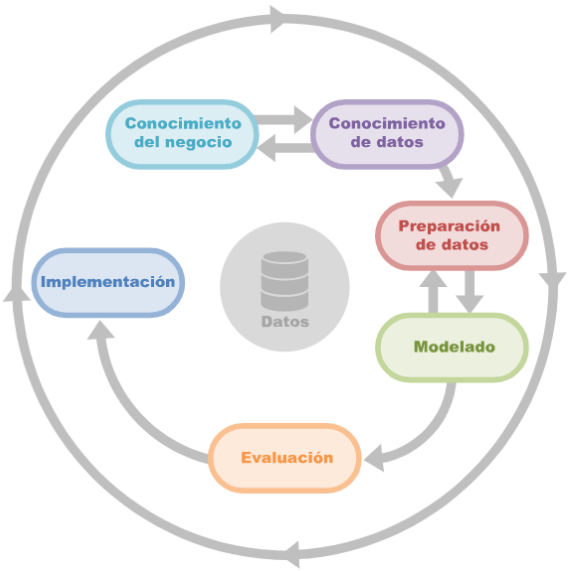


Ilustración 1 - Diagrama de flujo de la Metodología CRISP-DM

Al igual que las metodologías KDD y SEMMA, la sucesión de fases no es necesariamente rígida, el orden de las fases puede ser alterado, y se puede volver de una fase a la otra (Moine, Haedo , & Gordillo, 2011).

A continuación, se pasará a explicar más en detalle cada una de las fases de esta metodología (Chapman, y otros, 2000):

1- Comprensión del negocio: La fase inicial se centra en comprender los objetivos y requisitos del proyecto desde una perspectiva empresarial, para lograr transformarlo en un problema de minería de datos con sentido de negocio y crear un plan diseñado para lograr los objetivos. En esta fase se deben seleccionar las tecnologías y herramientas a utilizar y definir planes detallados para cada fase del proyecto.
2- Comprensión de los datos: La fase de comprensión de datos se centra en identificar, recopilar y analizar los conjuntos de datos que pueden ayudar a lograr los objetivos del proyecto. Esta fase tiene 4 tareas básicas. En primer lugar, se deben recopilar los datos. En segundo lugar, hay que examinar los datos para comenzar a familiarizarse con ellos identificando propiedades superficiales, como el formato de los datos, el número de registros o las identidades de los campos. En tercer lugar, se debe explorar los datos de manera más profunda logrando identificar relaciones entre los datos y detectando subconjuntos interesantes para formar hipótesis sobre información oculta. Por último, se deben identificar posibles problemas de calidad de los datos.
3- Preparación de los datos: La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto de datos final a partir de los datos brutos iniciales. Es probable que las tareas de preparación de datos se realicen varias veces y no en cualquier orden prescrito. Las tareas incluyen la selección del conjunto de datos a utilizar (documentando los motivos de inclusión/exclusión), limpieza de datos para herramientas de modelado, construcción y transformación de datos (a partir de datos que ya tengo) y formateo de los datos.
4- Modelado: En esta fase, se seleccionan y aplican varias técnicas de modelado, y sus parámetros se calibran para optimizar valores. Normalmente, existen varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requisitos sobre la forma de los datos. Por lo tanto, a menudo es necesario volver a la fase de preparación de datos. Esta fase tiene cuatro tareas. Seleccionar las técnicas de modelado. Principalmente en los algoritmos supervisados, se deben dividir los datos en conjuntos de entrenamiento, validación y prueba. Construir el modelo. Evaluar la calidad del modelo (se evalúa técnicamente).
5- Evaluación: La fase de evaluación analiza qué modelo se adapta mejor al negocio y se analiza qué hacer a continuación. Esta fase tiene tres tareas. En primer lugar, se evalúan los resultados: ¿Los modelos cumplen los criterios de éxito empresarial? ¿Cuáles debemos utilizar para el negocio? En segundo lugar, se revisa el trabajo realizado. ¿Se pasó algo por alto? ¿Se ejecutaron correctamente todos los pasos? Se deben resumir los hallazgos y corregir cualquier cosa si es necesario. Por último, se debe determinar si continuar con la implementación, repetir más modelos o iniciar nuevos proyectos.
6- Implementación: La creación del modelo generalmente no es el final del proyecto. Incluso si el propósito del modelo es aumentar el conocimiento de los datos, el conocimiento adquirido deberá organizarse y presentarse de manera que pueda ser utilizado. Dependiendo de los requisitos, la fase de

implementación puede ser tan simple como generar un informe o tan compleja como implementar un proceso de minería de datos repetible en toda la empresa. En muchos casos, es el cliente, no el analista de datos, quien lleva a cabo la implementación. Sin embargo, incluso si el analista llevara a cabo el esfuerzo de implementación, es importante que el cliente comprenda de antemano qué acciones deben llevarse a cabo para poder hacer uso de los modelos creados.

3.2. Segmentación de clientes

Si tengo claro el cliente que quiero encontrar, después es más fácil lograr los objetivos. Para lograr encontrar el cliente correcto, es clave lograr una buena segmentación, es una parte esencial de la estrategia de la empresa y especialmente, de un marketing eficiente. Se define la segmentación de clientes como el proceso de dividir a los clientes en grupos con necesidades, actitudes, actividades y comportamientos similares con el objetivo de conocerlos mejor y atender sus necesidades. Una vez creados los grupos de clientes, se debe elegir aquel o aquellos que resulten más apropiados para ser atendidos por la empresa (Peter & Olson, 2006).

Esta práctica es de vital importancia en la empresa ya que, con los resultados que se obtienen de la segmentación de clientes, se pueden identificar características claves del consumidor ideal. Una vez conocido el consumidor ideal, se puede saber cuál es el público objetivo en el cuál focalizarse. (Corrales, 2020)

Las segmentaciones estratégicas deben de ser generadas con sentido de negocio. Hay que intentar determinar las tipologías de clientes existentes en la empresa para determinar el tratamiento a aplicar en cada segmento. De esta forma se tratará de lograr los objetivos financieros de la empresa mediante la utilización de la ciencia de datos. En lugar de intentar llegar al 100% de los clientes, es más eficiente centrarse en un grupo específico de clientes que van a resultar siendo los más rentables para el negocio.

Existen muchos métodos alternativos para segmentar los clientes. Muchos de estos enfoques se derivan del campo del comportamiento del consumidor ya que la toma de decisiones de los clientes se ve afectada por factores racionales y emocionales (demografía, geografía, beneficios, motivaciones / necesidades, hábitos de compra, etc.). (Weinstein, 2004)

En el presente proyecto, se va a intentar segmentar a los clientes de diferentes maneras para lograr conocerlos y lograr aplicar las campañas de marketing de la forma más eficiente posible. En primer lugar, se segmentará a los clientes mediante la realización de análisis RFM. En segundo lugar, se segmentará a los clientes en base a su ingreso, cantidad consumida y antigüedad con la utilización de algoritmos de clustering. Por último, se intentará conocer el perfil del consumidor de cada uno de los productos incluidos en la base de datos, teniendo en cuenta las dos segmentaciones anteriores, con la utilización del algoritmo a priori.

3.2.1. Algoritmos de Clustering

Como fue mencionado anteriormente, la minería de datos es el acto de extraer información valiosa y conocimiento de grandes volúmenes de datos. La minería de datos se puede clasificar en dos grandes categorías: descriptivas y predictivas. Clustering es un ejemplo de metodologías descriptivas. (Parvaneh, Abbasimehr, & Tarokh, 2012)

Se define clustering como el conjunto de técnicas descriptivas que tienen por objetivo formar grupos a partir de un conjunto de datos. Los objetos dentro de cada clúster deben ser homogéneos y los clústeres deben ser heterogéneos entre sí. Asimismo, cada objeto debe pertenecer únicamente a 1 clúster. (Esteban, 2020)

Cabe destacar que, mientras Clustering es el proceso técnico de agrupamiento no supervisado, segmentación es el acto de crear segmentos de clientes o mercados. En este sentido y como fue mencionado anteriormente, se van a utilizar técnicas de clustering para segmentar a los clientes. (Ansari, 2021)

Los algoritmos de clustering pueden ser clasificados en las siguientes categorías: Métodos de Particionamiento, Métodos Jerárquicos, Métodos basados en Densidad y Métodos Basados en Cuadrícula. (Ansari, 2021) Se probará cuál de estos métodos es el que mejor aplica a nuestro conjunto de datos.

En el cuadro a continuación se presenta un breve resumen de cada uno de estos cuatro métodos:

Métodos de Clustering	Descripción	Algoritmos más comúnmente utilizados
Particionamiento	Divide los datos en k grupos, siendo k un numero especificado por el usuario. Cada grupo debe contener al menos un objeto y cada objeto debe pertenecer a un grupo. La mayoría de los métodos de particionamiento se basan en la distancia. No funcionan bien para clústeres con forma irregular.	K-medias, k-medoids, CLARANS, Algoritmo esperanza-maximización (EM)
Jerárquicos	Descompone los datos en una jerarquía de grupos, cada grupo más grande contiene un conjunto de subgrupos. Dos métodos: **aglomerativo** (también llamado enfoque ascendente), se parte de tantos grupos como objetos haya, y se van agrupando hasta que todos los grupos se convierten en uno; o **divisivo** (también llamado de arriba hacia abajo), comienza con todos los objetos en un mismo grupo grande y luego separa hasta que finalmente cada objeto está en un clúster. Pueden basarse en la distancia o la densidad y la continuidad	BIRCH (balanced iterative reducing and clustering using hierarchies), Chameleon, Método de Ward, vecino más cercano, dendrograma para visualización gráfica de la jerarquía
Basados en Densidad	Útil para clústeres con forma irregular. Su idea general es seguir creciendo un clúster determinado siempre y cuando la densidad (número de objetos o puntos de datos) en el "vecindario" supere algún umbral.	Distancia definida (DBSCAN), Escala múltiple (OPTICS), clúster basado en densidad (DENCLUE), SNN
Basados en Cuadricula	Cuantifican el espacio de objetos en un número finito de celdas que forman una estructura de cuadrícula. Todas las operaciones de agrupación en clústeres se realizan en la estructura de cuadrícula (es decir, en el espacio cuantificado).	STING (statistical information grid), Cluster Wave, clustering in quest (Clique).

Tabla 1 - Resumen explicativo de los distintos métodos de clustering (Ansari, 2021)

En una primera instancia, se intentará formar los clústeres con el algoritmo k-media. Si los clústeres formados resultantes son relativamente circulares y están bien definidos, entonces se podría decir que es correcto aplicar k-media para realizar los clústeres. En caso contrario, se probará algún otro tipo de algoritmo (métodos jerárquicos o de densidad).

Para corroborar que la utilización del algoritmo k-media para segmentar a los clientes del conjunto de datos es correcto, se harán dos cosas:

Primero, se pondrá a prueba el estadístico de Hopkins. Este estadístico permite evaluar la tendencia de clustering de un conjunto de datos mediante el cálculo de la probabilidad

de que dichos datos procedan de una distribución uniforme. (Kassambara, Data Novia, 2018)

El estadístico Hopkins busca para cada punto de mi base de datos, su vecino más cercano. Luego calcula la distancia que hay entre el punto y su vecino más cercano (Xi). Hace lo mismo para una base de datos aleatoria, denominando a la distancia desde el punto a su vecino más cercano como Yi. Luego se calcula el estadístico Hopkins como la sumatoria de las distancias de los puntos a sus vecinos más cercanos de la base de datos aleatoria, dividido entre la sumatoria de las distancias de la base de datos aleatoria más la sumatoria de las distancias de mi base de datos. Se puede observar en la siguiente ecuación:

Ecuación 1 - Estadístico de Hopkins

$$H = = \frac{\sum_{i=1}^{n} yi}{\sum_{i=1}^{n} xi + \sum_{i=1}^{n} yi}$$

Valores de H en torno a 0.5 indica que los datos se distribuyen uniformemente, por lo que no tendría sentido aplicar algoritmos de clustering. Cuanto más cercano a 1 sea H, mayores son las posibilidades que si se aplica un algoritmo de clustering correctamente, los clústeres resultantes sean buenos.

Una vez que confirmemos que los datos no están distribuidos uniformemente y por lo tanto hay tendencia de clustering, entonces se pasará a evaluar si aplicar el algoritmo k-media sería lo correcto. Por lo tanto, en segundo lugar, se realizará un análisis visual para identificar la distribución de los datos. Para poder observar los datos de forma gráfica, será necesario realizar un análisis de componentes principales de forma de reducir la dimensionalidad a únicamente 2 dimensiones, y graficar las observaciones de la dimensión 1 en función de la dimensión 2. Tanto en el objetivo secundario 3 como en el objetivo secundario 4, se busca segmentar a los clientes teniendo en consideración 3 variables. En este sentido, dos variables deberán ser “combinadas” en un componente.

De forma sintética, el método de componentes principales tiene como objetivo transformar un conjunto de variables, denominadas como *variables originales interrelacionadas*, en un nuevo conjunto de variables, combinación lineal de las originales, denominadas *componentes principales*. Estas últimas se caracterizan por estar correlacionadas entre sí. (Pérez López, 2004)

Por otra parte, para que los algoritmos de Clustering funcionen adecuadamente (especialmente el algoritmo k-medias), es importante que las variables estén normalizadas. Todas las variables a utilizar deben tener el mismo rango de valores.

Asimismo, es importante determinar el número óptimo de clústeres. Hay distintas cosas que se deben tener en cuenta a la hora de determinar la cantidad de clústeres óptima. En primer lugar, siempre se debe tener presente el sentido de negocio (para qué y por qué estoy realizando la segmentación). En este sentido, la cantidad de clústeres debe lograr segmentar a los clientes de forma que los clústeres formados brinden información relevante.

Una vez dicho esto, en segundo lugar, hay muchos métodos distintos para determinar el número de clústeres óptimo más allá del sentido de negocio (más de 30). Los dos más nombrados son los siguientes (Martínez R. G., 2021):

- Método Elbow: Examina la varianza total intra-cluster (WSS) como una función del número de clústeres. Se debe elegir el número de clústeres tal que agregar otro clúster no mejore mucho el WSS total. La idea detrás de los métodos de particionamiento es definir clústeres de modo que se minimice la variación total dentro del clúster. El método Elbow calcula la varianza total intra-cluster en función del número de clústeres, y escoge como óptimo aquel valor a partir del cual añadir más clústeres apenas consigue mejoría.
  El número de clústeres se puede definir de la siguiente manera:
  1- Ejecutar el algoritmo de clustering (por ejemplo, k-means clustering) para diferentes valores de k.
  2- Para cada k, calcular la variación total dentro del clúster (wss).
  3- Dibujar la curva de wss según el número de clústeres k.
  4- La ubicación de una curva (elbow) en la gráfica se considera generalmente como un indicador del número adecuado de clústeres.
- Método Average Silhouette: Este enfoque mide que tan buena es la asignación que se ha hecho de una observación comparando su similitud con el resto de las observaciones de su clúster, frente a las de otros clústeres. Es decir, determina la cantidad de la asignación de cada objeto dentro de su clúster. Un valor alto para el Average Silhouette indica una buena agrupación en clústeres.
  Se calcula utilizando la distancia media intra-clúster “a” y la distancia media más cercana al clúster “b” para cada muestra. El coeficiente de silueta para una muestra es (b – a) / max (a,b).
  El número óptimo de clústeres k es el que maximiza la silueta media en una gama de valores posibles para k. La silueta puede estar entre -1 y 1, siendo valores altos un indicativo de que la observación se ha asignado al clúster correcto.
  Este algoritmo es similar a elbow y se puede calcular de la siguiente manera:
  1- Ejecutar el algoritmo de clustering (por ejemplo, k-means clustering) para diferentes valores de k.
  2- Para cada k, calcular la silueta media de las observaciones.
  3- Dibujar la curva de avg. sil según el número de clústeres k.
  4- El punto en donde la curva toma el valor máximo se considera como el número adecuado de clústeres.

En el presente proyecto se corroborará no sólo la cantidad de clústeres óptima por estos dos métodos, sino que también se utilizará la librería *Nbclust* de R, que cuenta con una función que permite calcular el número óptimo de clústeres por 30 métodos distintos. (Charrad, Ghazzali, Boiteau, & Niknafs, 2014)

3.2.2. Algoritmo K-Medias

El algoritmo de clustering k-medias, es el algoritmo de Machine Learning no supervisado más comúnmente utilizado para particionar la base de datos en k grupos; donde k representa el número de grupos especificado por el usuario. Este algoritmo clasifica los objetos en múltiples grupos (clústeres) de forma que los objetos en un mismo clúster sean lo más parecidos posible entre sí y que los clústeres creados sean lo más diferentes posible. En los clústeres k-media, cada clúster está representado por su centro (centroide), que corresponde a la media de los objetos asignados al clúster. (Kassambara, 2017)

La idea básica detrás del algoritmo k-media, consiste en generar los k clústeres de la manera más compacta posible y que cada uno de los clústeres estén lo más separados posible. La calidad del clúster formado puede ser medida calculando la varianza interna (variación total dentro del clúster). La misma debe ser lo más pequeña posible. La varianza interna es la suma cuadrada de los errores entre todos los objetos en el clúster y el centroide. En otras palabras, para cada objeto en cada clúster, la distancia desde el objeto hasta el centroide del clúster al que pertenece se eleva al cuadrado y las distancias de todos los objetos se suman. (Han, Kamber, & Pei, 2012)

Pasos a seguir para la creación de clústeres con k-media:

1- Indicar el número de clústeres (k) que se van a generar.
2- El algoritmo comienza seleccionando de forma aleatoria k observaciones de la base de datos para que sirvan inicialmente como centroides de los clústeres.
3- Todas las observaciones restantes son asignadas al centroide más cercano, donde “más cercano” es definido utilizando la distancia Euclídea entre la observación y la media del clúster.
4- Una vez asignadas las observaciones, el algoritmo calcula la nueva media de cada uno de los clústeres formados (se suman todas las observaciones del clúster y se divide entre la cantidad de observaciones que contenga el clúster).
5- Dada la nueva media, se chequea que las observaciones se encuentren en el clúster correcto, ya que podría pasar que una observación esté más cerca de otro clúster. De esta manera, todas las observaciones son recolocadas utilizando la nueva media como centroide.
6- Los pasos 4 y 5 se repiten iterativamente hasta que las observaciones dejen de cambiar de clúster. Como podría pasar que no converja nunca, por lo general se pone un máximo de 10 iteraciones para que el algoritmo frene.

En la siguiente figura se puede observar el funcionamiento del algoritmo k-media:

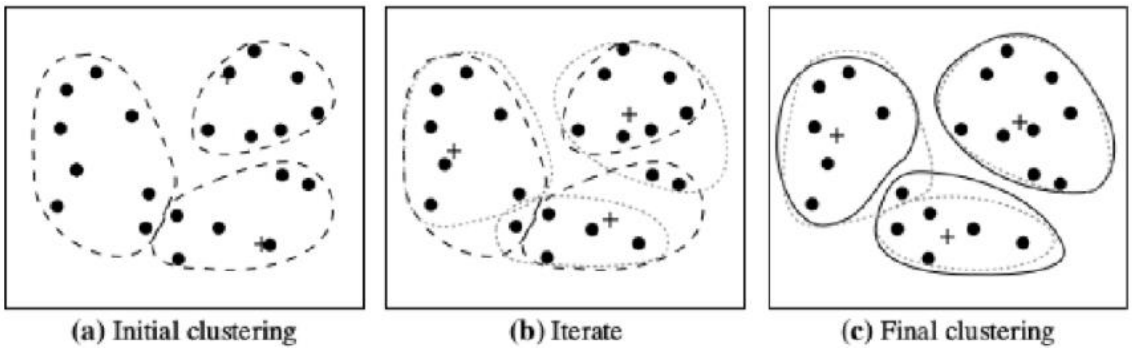


Ilustración 2 - Funcionamiento algoritmo k-media (Han, Kamber, & Pei, 2012)

El algoritmo k-media es un algoritmo muy fácil y rápido de utilizar y funciona de forma muy eficiente con grandes conjuntos de datos. De todas maneras, el algoritmo presenta algunas desventajas:

- Asume que los datos son conocidos y requiere que el usuario elija de antemano la cantidad de clústeres que se quieren formar. Esta desventaja lleva a que se tengan que probar diferentes números de clúster para encontrar el que mejor se adapte a los datos.

- El resultado final es sensible a la selección aleatoria inicial de los centroides. Esto es un problema, ya que cada vez que se corra el algoritmo, los clústeres pueden ser formados de manera diferente. Una forma de solucionar esto, es obligando al algoritmo a probar iniciar muchas veces con diferentes observaciones como centroide, para encontrar la que dé mejores resultados.
- Es sensible a datos atípicos. Esto se puede solucionar tratando los datos atípicos en la etapa de depuración de datos, o utilizando k-mediode en lugar de k-media.

3.2.3. Modelo RFM

El modelo RFM es una forma clara y descriptiva de clasificar a los clientes en función del comportamiento de compra. Esta herramienta ha sido utilizada con mucho éxito por los especialistas de marketing hace ya casi 20 años. A pesar de que últimamente se ha puesto más de moda la utilización de modelos predictivos e inteligencia artificial, el modelo RFM sigue siendo de gran utilidad en el marketing de base de datos moderno. Es especialmente útil cuando necesitamos definir estrategias para generar compromiso y lealtad en nuestra base de datos (Martinez R. G., Carrasco, Garcia-Madariaga, Porcel Gallego, & Herrera-Viedma, 2020).

El modelo de recencia, frecuencia y valor monetario (RFM), está relacionado con la adquisición, retención y gestión de relaciones de los clientes más rentables de la empresa. En este sentido, una vez conocido el modelo RFM, es importante que los departamentos de marketing sean capaces de realizar campañas de marketing más eficientes. (Martinez R., Carrasco, Sanchez-Figueroa, & Gavilan, 2021)

Asimismo, la razón por la cual el modelo RFM se sigue utilizando hasta el día de hoy es que "*es uno de los métodos de segmentación de clientes más sencillos de implantar, y al mismo tiempo uno de los que mejores resultados aportan a corto plazo. Se basa en el célebre principio de Pareto, según la cual el 20% de los clientes de una compañía generan el 80% de los ingresos"* (Córdoba, 2011). En este sentido, es importante lograr identificar quienes son ese 20% de los clientes para prestarles mayor atención. (Kumar & Reinartz, 2018)

Ilustración 3 - Ley de Pareto (el 20% de los clientes de una empresa generan el 80% de los ingresos) (Córdoba, 2011)

El modelo RFM es un modelo para determinar el valor del cliente en base a tres dimensiones:

Recencia (*Recency*): representa el tiempo que pasó desde la última compra del cliente.

Frecuencia (*Frequency*): representa cuantas veces el cliente compró en el periodo analizado.

Valor monetario (*Monetary*): representa el total del dinero gastado por el cliente en la empresa en el periodo de tiempo analizado.

Una vez obtenidas la recencia, frecuencia y valor monetario de cada uno de los clientes, el Modelo RFM lo que hace es asignar un puntaje del 1 al 5 en cada una de las tres dimensiones. En este sentido, se toman los datos de todos los clientes y se dividen en percentiles iguales (habitualmente quintiles). El 20% de los clientes con menor recencia van a obtener un puntaje 5, el 20% siguiente un 4, y así hasta llegar al último quintil de clientes correspondiente al 20% de los clientes con recencia más baja. Lo mismo se repite para las tres dimensiones. Una vez obtenido el puntaje para cada cliente en cada una de las 3 dimensiones, el modelo RFM le asigna el puntaje global multiplicando cada uno de los puntajes obtenidos en las dimensiones por *W(R,F,M)*. Siendo el puntaje global: *RFMScore = RecencyScore × wR + FrequencyScore × wF + MonetaryScore × wM.* W puede ser = 1/3, asignándole el mismo peso a cada una de las tres dimensiones, o puede variar dependiendo de si se le quiere dar más importancia a alguna de las tres dimensiones. El peso que se le da a cada una de las tres dimensiones del modelo RFM, es un parámetro a estudiar. En las empresas que funcionan en base a suscriciones mensuales, por ejemplo, la recencia es sumamente importante. Por otra parte, en tiendas como Ikea, tal vez la recencia no es la dimensión más importante del modelo. De todas maneras, la suma de todos los pesos tiene que ser siempre igual a 1. De esta forma, una vez asignados los pesos y multiplicado por cada una de las dimensiones, se les da a los clientes un *"valor*". El valor puede variar entre 1 y 5, siendo 5 el cliente con mayor valor y 1 el cliente con menor valor.

En marketing, el valor del cliente en el tiempo (CLV) es el valor que el cliente contribuye en el ciclo de vida de una empresa. Es una métrica muy útil, ampliamente usada por los gerentes de marketing, especialmente cuando se quieren centrar en la adquisición y retención de clientes. Como fue mencionado anteriormente, el RFM score es una de las tantas maneras de calcular el valor del cliente (CLV). (Martinez R. , Carrasco, Sanchez-Figueroa, & Gavilan, 2021)

De todas maneras, el RFM score como tal es de utilidad muy pocas veces, ya que hay 125 combinaciones posibles para asignar el score global. Un puntaje de 4 puede ser un cliente nuevo que realizó una única compra grande (con frecuencia baja, pero valor monetario y recencia alta), puede ser un cliente con recencia y frecuencia alta y valor monetario bajo, o un cliente que está a punto de abandonar la empresa con recencia baja, pero valor monetario y frecuencia alta, etc.

Considerar cada una de las dimensiones por separado puede ser mucho más útil, pudiendo generar clústeres de clientes, dependiendo del puntaje obtenido en cada una de las tres dimensiones. (Wright, 2021)

En este sentido, aparte del valor del cliente, es importante conocer el ciclo de vida del cliente. El mismo se podría dividir en 6 etapas: conocimiento, adquisición, conversión, crecimiento, retención y reactivación. El modelo RFM en cierta medida me brinda información sobre en cuál de las 6 etapas se encuentra el cliente. Es importante conocer esto, ya que la campaña de marketing a aplicar en cada una de las etapas es diferente.

Un cliente con recencia, frecuencia y valor monetario alto serían los clientes VIP, de mayor valor para la empresa. Un cliente con Recencia alta y frecuencia baja podría ser

considerado nuevo cliente, teniéndole que aplicarles campañas de adquisición. Un cliente con recencia baja, pero valor monetario alto, podría ser un cliente que está a punto de abandonar la empresa, teniéndole que aplicar campañas de retención.

En la ilustración 2 se puede ver un ejemplo de un cliente con RFM (5, 4, 2). Es un cliente que se encuentra entre el 20% de los clientes más recientes, que está en el segundo quintil de frecuencia, pero tiene un valor monetario relativamente bajo, ya que se encuentra entre el segundo quintil más bajo en cuanto al valor monetario. Podríamos estar hablando de un cliente que se incorporó recientemente a la empresa:

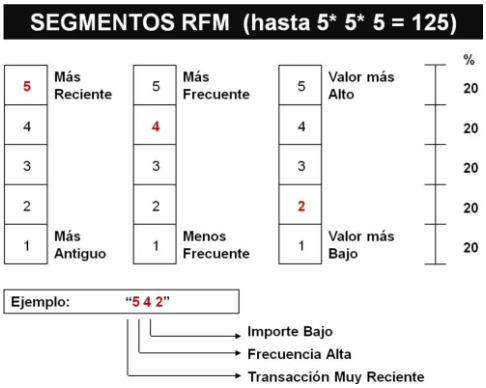


Ilustración 4 - Ejemplo RFM Score (Córdoba, 2011)

De todas maneras, el modelo RFM clásico tiene algunas debilidades principalmente asociada a la falta de precisión. Al utilizar una escala del 1 al 5 considerando cada una de las dimensiones de forma independiente, no se está teniendo en cuenta la posible correlación que hay en cada una de las tres dimensiones.

El modelo RFM no brinda información precisa acerca de si un cliente es nuevo. Se tiende a pensar que un cliente con Recencia alta y Frecuencia baja es un cliente nuevo, pero se ha comprobado que podría corresponder a un cliente con fecha de registro muy antigua, pero que ha vuelto a comprar este año después de mucho tiempo. En este caso, el cliente no sería realmente nuevo.

En el presente proyecto, se realizarán clústeres de clientes con las tres dimensiones normalizadas, para encontrar los hábitos de compra de los clientes. Se van a intentar identificar al menos los siguientes segmentos:

- Vip: son los mejores clientes de la empresa (recencia, valor monetario y frecuencia alta)
- Churn: aquellos clientes que solían ser buenos clientes y por una cosa u otra están abandonando a la empresa (valor monetario y frecuencia media / alta, pero recencia baja)
- Nuevos: los nuevos clientes (o que se están reactivando) que parecerían ser prometedores (valor monetario y frecuencia media o baja, pero recencia alta)
- Peores: aquellos que no deberían ser considerados clientes, ya que su hábito de compra es prácticamente nulo (valor monetario, frecuencia y recencia baja).

Dependiendo de los resultados, tal vez se podrá identificar algún otro segmento.

3.3. Algoritmo a priori

El algoritmo a priori fue propuesto en 1994 por Agrawal y Srikant. Fue uno de los primeros algoritmos desarrollados para la búsqueda de reglas de asociación y sigue siendo uno de los más empleados (Naranjo Cuervo & Sierra Martínez, 2009). Antes de comenzar con la explicación sobre cómo funciona el algoritmo, me gustaría responder la siguiente pregunta:

¿Qué son reglas de asociación? Las reglas de asociación buscan relaciones recurrentes dentro de un conjunto de datos determinado. Tienen como objetivo descubrir asociaciones y correlaciones interesantes entre elementos de una base de datos transaccional. Un ejemplo comúnmente utilizado en reglas de asociación es "el análisis de canasta de mercado". Analiza los hábitos de compra de los clientes mediante la búsqueda de asociaciones entre los diferentes artículos que los clientes colocan en sus "cestas de compra". En este caso, los analistas del mercado pueden utilizar la información para agrupar productos físicamente en la tienda para aumentar las posibilidades de venta cruzada, impulsar los motores de recomendación online, dirigir campañas de marketing mediante el envío de cupones promocionales a los clientes para ofrecer productos relacionados con artículos que compraron recientemente, etc. (Martinez R. G., 2021)

Algunos conceptos: Cada uno de los elementos que forman parte de determinada transacción, son conocidos como *item* y cada conjunto de ellos *itemset.* Las transacciones pueden estar formadas por uno o varios *items.* En el caso que este formada por varios, cada subconjunto de ellos es un *itemset* distinto. A modo de ejemplo:

La transacción T = {A, B, C} está formada por 3 items (A, B y C) y sus posibles *itemsets* son: {A, B, C}, {A, B}, {B, C}, {A, C}, {A}, {B}, y {C}.

Una regla de asociación supone que "si pasa X, entonces pasa Y" (X->Y), donde X e Y son *itemsets* o *items* individuales. En este caso, la X (lado izquierdo) es el antecedente y la Y (lado derecho) es el consecuente. (Rodrigo, Ciencia de Datos, 2018). Más conceptos:

- Soporte: Es el porcentaje de transacciones que contienen todos los elementos de un *itemset*. Es el número de transacciones que contienen el *itemset* X, dividido el total de transacciones. En una regla, el soporte es el porcentaje de transacciones que tienen todos los *items* de la regla (tanto el consecuente como el antecedente). Es la frecuencia relativa del itemset. Se busca que los valores del soporte san lo más altos posibles.
- Confianza: Es la probabilidad de que una transacción que contiene elementos del antecedente (lado izquierdo de la regla), también contiene el artículo del consecuente (lado derecho de la regla), y se define con la siguiente ecuación:

Ecuación 2 - Confianza itemset

$$Confianza(X \Longrightarrow Y) = \frac{soporte\big(\acute{u}nion(X,Y)\big)}{soporte(X)}$$

Donde unión (X, Y) es el itemset que contiene todos los items de X y de Y.

Es entonces la probabilidad empírica de que ocurra el consecuente dado que ocurrió el antecedente. Valores de confianza altos, significa que las probabilidades de que

si sucedió X también suceda Y, son altas. La confianza tiene como objetivo medir la calidad de la predicción de la regla, basándose en que sucederá en el futuro a partir de lo que ya sucedió en transacciones anteriores.

- Lift: Este estadístico compara la frecuencia observada de una regla con la frecuencia esperada simplemente por azar. Refleja el aumento de la probabilidad de que ocurra el consecuente cuando nos enteramos de que ocurre el antecedente. El valor Lift de una regla sigue la siguiente ecuación:

Ecuación 3 - Lift itemset

$$Lift(X \Longrightarrow Y) = \frac{soporte\big(\acute{u}nion(X,Y)\big)}{soporte(X) * soporte(Y)}$$

Cuanto más lejos se encuentre el Lift de 1, mejor es la calidad de la regla. Es decir, existen más evidencias de que la regla no se haya dado simplemente por azar, lo que es lo mismo a decir que más evidencias de que la regla representa un patrón real.

El algoritmo a priori es uno de los tantos algoritmos diseñados para identificar itemsets frecuentes y reglas de asociación.

El algoritmo tiene dos etapas:

1- Identificar todos los *itemsets* que ocurren con una frecuencia por encima de un determinado límite (*itemsets* frecuentes).
2- Convertir esos *itemsets* frecuentes en reglas de asociación.

El "soporte" definido anteriormente, es el índice para la generación de *itemsets* y la "confianza" es el índice para la generación de reglas de asociación.

Encontrar itemsets frecuentes (*itemsets* con una frecuencia mayor o igual a un determinado soporte mínimo) conlleva a un proceso computacional muy grande debido a la cantidad de combinaciones posibles, sin embargo, una vez identificados, no es difícil generar reglas de asociación que tengan una confianza mínima. El algoritmo a priori realiza una búsqueda exhaustiva de *itemsets* comenzando con los *itemsets* de menor tamaño y terminando con los de mayor tamaño, pero realiza la búsqueda de forma optimizada.

Se observa la siguiente figura para explicar la idea principal del algoritmo:

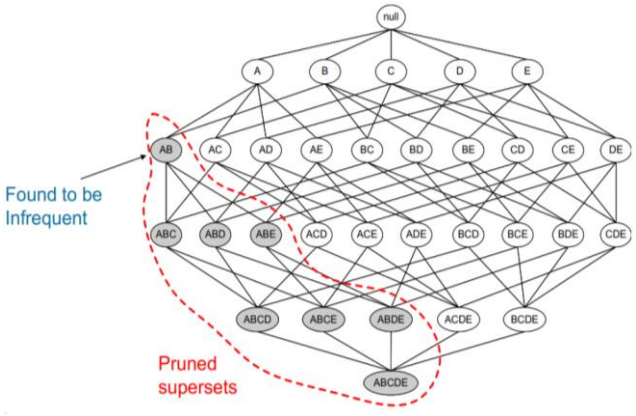


Ilustración 5 - Funcionamiento algoritmo a priori

En la imagen se encuentran los productos (A, B, C, D y E) y todas las posibles combinaciones entre ellos. Para reducir la carga computacional que implicaría buscar reglas de asociación entre todos los productos de una empresa, lo que se utiliza es la regla anti monótona para la poda (Pitol, 2014). El algoritmo a priori intenta optimizar la búsqueda de reglas, simplificando lo más posible todas las posibles combinaciones. Si un conjunto es infrecuente, entonces todos los conjuntos en donde este último se encuentre también serán infrecuentes. Volviendo a la imagen, a lo que se refiere el algoritmo es que, si {AB} **no** es frecuente, entonces todos los *itemsets* que incluyan el *itemset* {AB}, ({ABC}, {ABE}, {ADBE}, etc.) también serán infrecuentes, por el simple hecho de contener a A y B en ellos.

Entonces, el algoritmo a priori comienza con los *items* individuales que tengan frecuencia suficiente y sigue una estrategia *bottom-up*, en donde va agregando *items* al *itemset* de a uno, eliminando aquellos subconjuntos que no alcancen el soporte mínimo (Rodrigo, Ciencia de Datos, 2018). El soporte y la confianza mínima necesaria es estipulada por el usuario.

En el presente proyecto no se utilizará el algoritmo a priori para obtener “canastas de compra”, ya que tenemos únicamente 6 productos en la base de datos. El algoritmo a priori se utilizará para encontrar el perfil del cliente que consume cada uno de esos 6 productos. Para ello, se tendrán en cuenta tanto las características personales de cada cliente, como los clústeres en los que se encuentran, generados en etapas previas.

4. Desarrollo del trabajo y principales resultados

En este capítulo se realizará el proceso completo CRISP-DM aplicado a la empresa mencionada anteriormente, para lograr cumplir con los objetivos propuestos. Para su desarrollo se utilizarán las herramientas KNIME, SAS Miner y R. Tanto, la configuración de los nodos utilizados en KNIME y SAS Miner como el código utilizado en R se encuentran en el anexo.

4.1. Comprensión del Negocio

La base de datos a utilizar en el proyecto fue extraída de la página Kaggle (Saldanha, 2021). El objetivo propuesto en el concurso de Kaggle es predecir quién responderá a una oferta de un producto o servicio. Plantean un problema de aprendizaje supervisado con la variable *Response* (que es 1 si el cliente respondió a la oferta en la última campaña y 0 en caso contrario) como variable objetivo.

Para el presente proyecto, se va a utilizar la base de datos para un problema distinto que el propuesto en el concurso. Toda empresa tiene clientes de diferentes características y con grados de importancia diferentes. Para que la empresa logre conocer cuáles son los clientes que realmente son importantes, es necesario realizar una correcta segmentación. En este sentido, la base de datos será utilizada para estudiar los perfiles de los clientes, identificando cuáles son los clientes más valiosos para la empresa y cuáles serían más propensos a realizar tipos de compras determinadas.

Los datos fueron extraídos de una empresa bien establecida que opera en el sector minorista de alimentos. Según descrito en la página de Kaggle, tienen alrededor de varios cientos de miles de clientes registrados y atienden a casi un millón de

consumidores al año. Venden productos de 6 categorías principales: vinos, productos cárnicos raros, frutas exóticas, pescados especialmente preparados, productos dulces y productos de bazar. Los clientes pueden pedir y adquirir productos a través de 3 canales de venta: tiendas físicas, catálogos y sitio web de la empresa. A nivel mundial, la compañía había tenido ingresos sólidos y un balance de resultados saludable en los últimos 3 años, pero las perspectivas de crecimiento de ganancias para los próximos 3 años no eran prometedoras. Por esta razón, se consideraron varias iniciativas estratégicas para revertir esta situación. Una de ellas fue intentar mejorar el desempeño de las actividades de marketing, con un enfoque especial en la eficiencia de campañas de marketing con la utilización de minería de datos.

4.2. Comprensión de los Datos

Como fue mencionado anteriormente, los datos fueron exportados de Kaggle. Los datos son de diciembre 2014. La base de datos tiene 2240 observaciones y las siguientes variables:

Variables	Tipo	Descripción
Age	Intervalo	Edad del cliente
Education	Clase	Nivel educativo del cliente
Mariteal_Status	Clase	Estado civil del cliente
Income	Intervalo	Ingreso familiar anual del cliente
Kidhome	Clase	Número de niños en el hogar del cliente
Teenhome	Clase	Número de adolescentes en el hogar del cliente
Antiquity	Intervalo	Número de días desde la primera compra del cliente
Recency	Intervalo	Número de días desde la última compra del cliente
MntWines	Intervalo	Cantidad gastada en productos vitivinícolas en los últimos dos años
MntFruits	Intervalo	Cantidad gastada en frutas en los últimos dos años
MntMeatProducts	Intervalo	Cantidad gastada en productos de carne en los últimos dos años
MntFishProducts	Intervalo	Cantidad gastada en productos de pescado en los últimos dos años
MntSweetProducts	Intervalo	Cantidad gastada en productos dulces en los últimos dos años
MntGoldProds	Intervalo	Cantidad gastada en productos de oro en los últimos dos años
NumDealsPurchases	Intervalo	Número de compras realizadas con descuento
NumWebPurchases	Intervalo	Número de compras realizadas a través de del sitio web de la empresa
NumCatalogPurchases	Intervalo	Número de compras realizadas mediante catalogo
NumStorePurchases	Intervalo	Número de compras realizadas directamente en tiendas
NumWebVisitsMonth	Intervalo	Número de visitas al sitio web de la empresa en el último mes
Complain	Binomial	1 si el cliente se quejó en los últimos 2 años, 0 en caso contrario
AcceptedCmp1	Binomial	1 si el cliente aceptó la oferta en la primera campaña, 0 en caso contrario
AcceptedCmp2	Binomial	1 si el cliente aceptó la oferta en la segunda campaña, 0 en caso contrario
AcceptedCmp3	Binomial	1 si el cliente aceptó la oferta en la tercera campaña, 0 en caso contrario
AcceptedCmp4	Binomial	1 si el cliente aceptó la oferta en la cuarta campaña, 0 en caso contrario
AcceptedCmp5	Binomial	1 si el cliente aceptó la oferta en la quinta campaña, 0 en caso contrario
Response	Binomial	1 si el cliente aceptó la oferta en la última campaña, 0 en caso contrario

Tabla 2 - Descripción de las posibles variables a utilizar

De esta manera se puede observar que contamos con 22 variables numéricas (15 de intervalo y 7 binomiales) y 4 variables de clase.

Cabe destacar que, en lugar de brindar la edad del cliente, brindan el año de nacimiento. Asimismo, *Antiquity* (antigüedad) fue dada en formato fecha. Ambas fueron modificadas en KNIME para representar la edad del cliente (en cantidad de años) y la cantidad de

días desde la primera compra del cliente. Para la creación de las variables *Age y Antiquity* fue tenido en consideración que los datos fueron extraídos en diciembre 2014.

Por otra parte, al exportar la base de datos de Kaggle, la misma incluye las variables *Z_Revenue y Z_CostContact.* Ambas son variables unarias por lo que no aportan ningún tipo de información. Fueron eliminadas de la base de datos.

Los datos fueron importados en SAS Miner con la utilización del asesor avanzado. De esta manera, es probable que variables numéricas sean considerados variables nominales.

Nombre	Rol	Nivel	Número de niveles	Porcentaje de ausentes	Mínimo	Máximo	Media	Desviación estándar
AcceptedCmp4	Input	Binario	2	0	.	.	.	.
AcceptedCmp2	Input	Binario	2	0	.	.	.	.
Complain	Input	Binario	2	0	.	.	.	.
AcceptedCmp5	Input	Binario	2	0	.	.	.	.
AcceptedCmp3	Input	Binario	2	0	.	.	.	.
AcceptedCmp1	Input	Binario	2	0	.	.	.	.
Response	Input	Binario	2	0	.	.	.	.
MntFruits	Input	Intervalo	.	0	0	199	26.30223	39.77343
MntWines	Input	Intervalo	.	0	0	1493	303.9357	336.5974
MntGoldProds	Input	Intervalo	.	0	0	362	44.02188	52.16744
MntFishProducts	Input	Intervalo	.	0	0	259	37.52545	54.62898
Income	Input	Intervalo	.	1.071429	1730	666666	52247.25	25173.08
Antiquity	Input	Intervalo	.	0	185	884	538.5821	202.1225
Age	Input	Intervalo	.	0	18	121	45.1942	11.98407
MntMeatProducts	Input	Intervalo	.	0	0	1725	166.95	225.7154
Recency	Input	Intervalo	.	0	0	99	49.10938	28.96245
MntSweetProduc	Input	Intervalo	.	0	0	263	27.06295	41.2805
ID	ID	Intervalo	.	0	0	11191	5592.16	3246.662
NumStorePurcha	Input	Nominal	14	0	.	.	.	.
NumCatalogPurc	Input	Nominal	14	0	.	.	.	.
NumWebPurchas	Input	Nominal	15	0	.	.	.	.
NumDealsPurcha	Input	Nominal	15	0	.	.	.	.
NumWebVisitsMo	Input	Nominal	16	0	.	.	.	.
Teenhome	Input	Nominal	3	0	.	.	.	.
Kidhome	Input	Nominal	3	0	.	.	.	.
Education	Input	Nominal	5	0	.	.	.	.
Marital_Status	Input	Nominal	8	0	.	.	.	.

Tabla 3 - Estadísticos asesor avanzado SAS Miner.

De la tabla anterior se puede decir lo siguiente:

- La edad mínima de los clientes es 18 años y la edad máxima de 121. La edad media es de 45 años.
- El ingreso promedio es de 52247.25. El ingreso más bajo es de 1725 y el ingreso más alto es de 666666, lo cual parecería ser un error.
- Los productos más consumidos son los vinos, con una media de 303.9357 dólares, seguido de la carne que tiene una media de consumo de 166.95. El resto de los productos tienen una media similar de consumo entre sí y están muy por debajo del consumo de productos vitivinícolas y carne. El consumo medio del resto de los productos ronda entre los 26.3 y los 44.02 dólares. Cabe destacar que esta diferencia también puede ser debida a diferencia en los precios de los productos. De todas maneras, se puede afirmar, que los clientes tienen un gran consumo de productos vitivinícolas.
- El cliente más nuevo lleva 185 días consumiendo en la empresa y el cliente más antiguo lleva 884 días consumiendo en la empresa. Los clientes tienen una antigüedad media de 538 días. Asimismo, el cliente que lleva más tiempo sin consumir en la empresa lleva 99 días sin consumir. Tienen una recencia media de 49 días.

A continuación, se presenta un gráfico de cajas con todos los productos consumidos por su frecuencia:

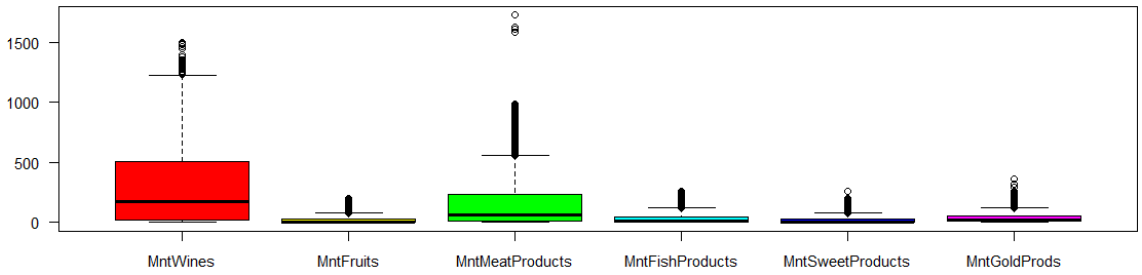


Ilustración 6 - Gráficos de caja de cantidad consumida de cada uno de los productos de la base de datos

De esta forma se puede observar de forma más evidente que los productos vitivinícolas son los más vendidos en la empresa, seguido de la carne. El consumo del resto de los productos es similar.

Las variables NumWebPurchases, NumWebVisitsMonth, NumCatalogPurchases, NumStorePurchases y NumDealsPurchases son variables numéricas que al utilizar el asesor avanzado de SAS Miner fueron consideradas variables nominales. De todas maneras, tienen demasiadas categorías, por lo que tal vez es mejor tomarlas como variables de intervalo. Si existe un orden entre las categorías (en este caso es evidente, ya que son el número de compras realizadas), es mejor usarlas como continuas. De todas maneras, se van a observar todas las variables gráficamente para ver que las categorías no presenten saltos muy bruscos:

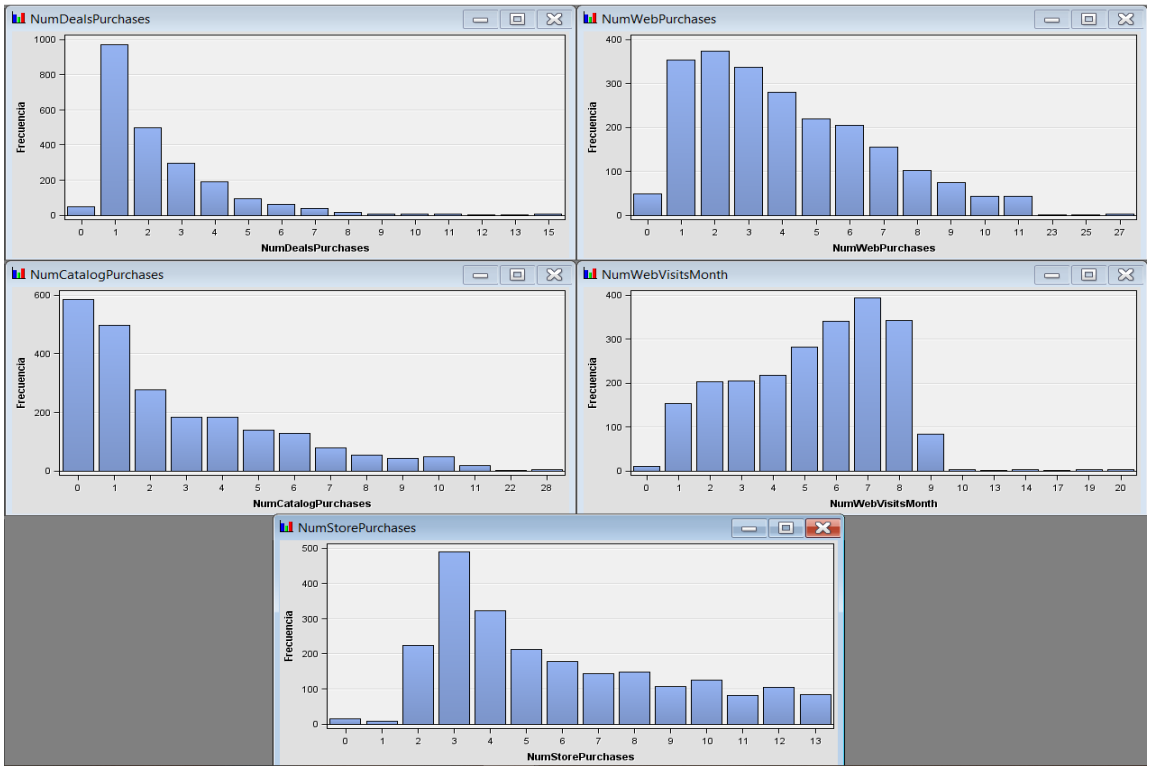


Ilustración 7 - Gráficos de barra de las variables relacionadas con el número de compras realizadas por los clientes

Ninguna de las variables realiza saltos bruscos. En este sentido, todas ellas serán tratadas como variables de intervalo, ya que, salvo excepciones, los algoritmos de Machine Learning suelen aprovechar mejor la información de esta manera.

Al observar las variables, veo que se pueden crear nuevas variables a partir de las ya existentes que podrían resultar útiles para comprender mejor el conjunto de datos y revelar información valiosa.

Las siguientes variables ya fueron creadas en KNIME en un paso previo a importar los datos en SAS Miner:

- ***Age**:* Edad del cliente. Reemplazó y se creó a partir de la variable *Year_Birth. Year_Birth* nos brinda únicamente información del año de nacimiento del cliente. De esta manera, se puede obtener la edad para obtener resultados más intuitivos.
- ***Antiquity**:* Número de días desde la primera compra del cliente. Reemplazó y se creó a partir de la variable *DtCustomer*. La variable *DtCustomer* es la fecha de la primera compra del cliente *(formato: aaaa-mm-dd)*. Al igual que con la variable edad, obtener la antigüedad del cliente en cantidad de días desde la primera compra puede arrojar resultados más intuitivos que trabajar con la fecha.

En la fase de preparación de los datos, se crearán las siguientes variables:

- ***Monetary**:* Suma del monto total gastado en las 6 categorías de productos.
- ***Frequency**:* Suma de número de compras realizadas en los 3 canales de venta: tiendas físicas, catálogos y sitio web de la empresa.
- ***Response**:* Será creada de una manera sutilmente diferente. El concurso de Kaggle propone la variable *Response* como 1 si el cliente aceptó la oferta en la sexta (y última) campaña de marketing, 0 en caso contrario. Yo voy a optar por crear la variable *Response* a partir de las seis campañas realizadas. En este sentido, será el promedio de respuesta de los clientes en las 6 campañas realizadas.

Una vez creadas las nuevas variables en KNIME, se hará una depuración de datos con la utilización de SAS Miner.

4.3. Preparación de los Datos

La preparación de los datos se va a dividir en dos etapas. En primer lugar, se crearán las variables necesarias en KNIME y luego se realizará una correcta depuración de datos con SAS Miner.

4.3.1. Creación de nuevas variables

Se van a crear las nuevas variables anteriormente mencionadas con la herramienta KNIME.

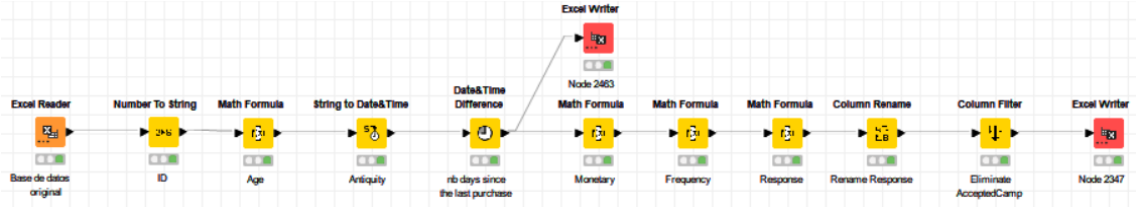


Ilustración 8 - Flujo creación nuevas variables en KNIME

4.3.2. Depuración de datos

Para comenzar con la depuración de datos, se observan las variables de intervalo con la utilización del nodo DMDB:

Variable	Ausente ▼	N	Mínimo	Máximo	Media	Desviación estándar	Asimetría	Curtosis
Income	24	2216	1730	666666	52247.25	25173.08	6.763487	159.6367
Age	0	2240	18	121	45.1942	11.98407	0.349944	0.717464
Antiquity	0	2240	185	884	538.5821	202.1225	-0.01522	-1.19465
Frequency	0	2240	0	32	12.53705	7.205741	0.297114	-1.11796
MntFishProducts	0	2240	0	259	37.52545	54.62898	1.919769	3.096461
MntFruits	0	2240	0	199	26.30223	39.77343	2.102063	4.050976
MntGoldProds	0	2240	0	362	44.02188	52.16744	1.886106	3.551709
MntMeatProducts	0	2240	0	1725	166.95	225.7154	2.083233	5.516724
MntSweetProdu...	0	2240	0	263	27.06295	41.2805	2.136081	4.376548
MntWines	0	2240	0	1493	303.9357	336.5974	1.175771	0.598744
Monetary	0	2240	5	2525	605.7982	602.2493	0.860841	-0.34194
NumCatalogPur...	0	2240	0	28	2.662054	2.923101	1.880989	8.047437
NumDealsPurch...	0	2240	0	15	2.325	1.932238	2.418569	8.936914
NumStorePurch...	0	2240	0	13	5.790179	3.250958	0.702237	-0.62205
NumWebPurcha...	0	2240	0	27	4.084821	2.778714	1.382794	5.703128
NumWebVisitsM...	0	2240	0	20	5.316518	2.426645	0.207926	1.821614
Recency	0	2240	0	99	49.10938	28.96245	-0.00199	-1.2019
Response	0	2240	0	66.66667	5.669643	11.68618	2.358081	5.821177

Tabla 4 - Estadísticos de sumarización variables de intervalo (nodo DMDB, SAS Miner)

Se puede observar que hay dos variables que pueden presentar errores. La edad, ya que marca una edad máxima de 121 años, lo cual es casi imposible, y el ingreso, que marca un ingreso máximo de 666666 lo cual aparenta ser un dato faltante. Asimismo, Income es la única variable con datos ausentes. Presenta 24 datos ausentes. Los mismos serán tratados más adelante. Es normal que variables como *Income*, presente datos ausentes, ya que es una información que a las personas no les apetece brindar. De todas maneras, podría ocurrir que el simple hecho que falte la información en esa variable, en realidad nos esté brindando información relevante.

Si se observa un gráfico de barras de cada una de las variables:

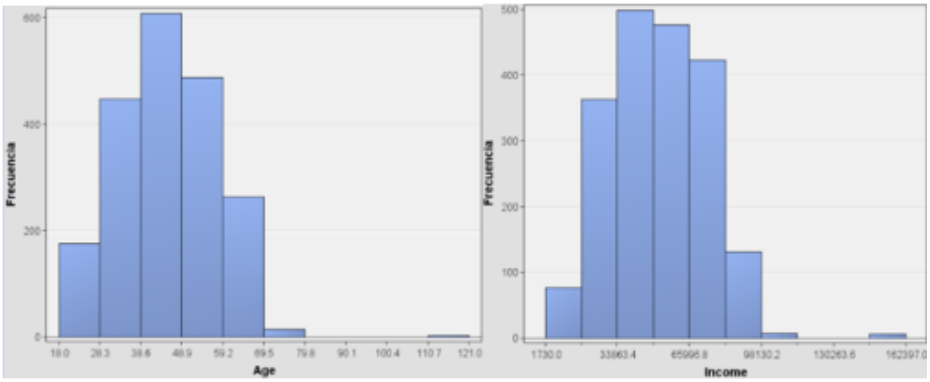


Ilustración 9 - Gráficos de barras variables Age e Income

En la variable Income se ve un salto que se pasa de ingresos de alrededor de 100000, a ingresos de entre 150000 y 162397. De todas maneras, estos datos no serán tratados ahora. En todo caso serán tratados más adelante como datos atípicos y no como errores.

Con la utilización del nodo reemplazo, se les va a poner a ambas variables un límite superior de reemplazo. Todas las observaciones que tengan una edad mayor a 100 y todas las observaciones que tengan un ingreso mayor a 666665 van a ser sustituidos por datos faltantes.

La base de datos presenta 4 variables de clase y 1 binarias (las variables AcceptedCmp fueron eliminadas luego de que la nueva variable “Response” fue creada). Ninguna de ellas presenta datos ausentes:

Nombre de la variable	Nivel	Número de ocurrencias	Porcentaje
Complain	0	2219	99.0625
Complain	1	21	0.9375
Education	Graduation	1127	50.3125
Education	PhD	486	21.69643
Education	Master	370	16.51786
Education	2n Cycle	203	9.0625
Education	Basic	54	2.410714
Kidhome	0	1293	57.72321
Kidhome	1	899	40.13393
Kidhome	2	48	2.142857
Marital Status	Married	864	38.57143
Marital Status	Together	580	25.89286
Marital Status	Single	480	21.42857
Marital Status	Divorced	232	10.35714
Marital Status	Widow	77	3.4375
Marital Status	Alone	3	0.133929
Marital Status	Absurd	2	0.089286
Marital Status	YOLO	2	0.089286
Teenhome	0	1158	51.69643
Teenhome	1	1030	45.98214
Teenhome	2	52	2.321429

Tabla 5 - Estadísticos de sumarización variables de clase (nodo explorador de estadísticos, SAS Miner)

Todas las categorías en amarillo no llegan a tener un 5% de observaciones. Por esta razón, con la utilización del nodo reemplazo van a ser reagrupadas con otras categorías.

La variable *complain* va a tener que ser rechazada, ya que al nivel 1 lo toman menos de un 5% de las observaciones y, por lo tanto, tiene solo 1 nivel válido.

Las variables *Kidhome* (número de niños en el hogar del cliente) y *Teenhome* (número de adolescentes en el hogar del cliente), tienen 3 niveles distintos (0, 1 o 2 niños/adolescentes). El nivel 2, en ninguna de las dos variables, llega a tener un 5% de las observaciones, por lo tanto, fueron reagrupados con el nivel 1. Ambas variables, pasaron a ser binarias, en donde toman el valor 0 si no hay niños/adolescentes en la casa, o 1 si los hay.

En la variable *Marital Status*, hay muchas categorías que no llegan a tener un 5% de las observaciones, por esta razón, se va a reagrupar todas las categorías en solamente dos, “*Single*” y “*In Couple*”.

En la variable *Education*, el nivel “*Basic”* tiene menos de un 5% de las observaciones. Este nivel fue reagrupado con el nivel “*2n Cycle”*.

Una vez aplicado el nodo reemplazo, con la utilización del nodo metadatos se va a rechazar a la variable complain.

4.3.2.1. Tratamiento de datos atípicos variables de intervalo:

La intención de este proyecto es trabajar con el algoritmo k-media. El mismo es sensible a datos atípicos. Por esta razón he decidido tratarlos de manera correcta. Las variables que van a ser utilizadas para la generación de clústeres con el algoritmo k-media son por un lado las tres dimensiones del modelo RFM (*Recency, Frequency y Monetary*) y por otro *Income, Antiquity y Monetary*. Se tratarán únicamente los atípicos de estas 5 variables. Se considerarán atípicos siempre y cuando la cantidad de datos “considerados como atípicos” sean menos de un 2% de las observaciones. Como para el presente proyecto se trabajará con Machine Learning no supervisado, no se van a dividir los datos en train y test. Se trabajará con un único conjunto de datos.

Con atípicos, me refiero a que se alejan mucho de la media o mediana (dependiendo de la asimetría de la variable). Estos datos pueden ser tratados de tres maneras distintas. Los datos simétricos son aquellos con un coeficiente de asimetría de entre -1 y 1 aproximadamente (su distribución es parecida a la de la normal). Para estos casos, los datos atípicos se tratarán con el método de desviación estándar (se considerarán

atípicos, aquellos datos que disten 3 veces de la desviación típica de la media). Los datos asimétricos son aquellos que están fuera del rango -1 y 1. Estos datos se tratarán, siempre y cuando la mediana sea distinta de cero, con el método de desviación absoluta, mediana (se tratarán como atípicos, todas aquellas observaciones que disten más de 9 medianas absolutas de la mediana). Si la mediana es 0, entonces los datos fuera del rango de entre -1 y 1 se tratan con el método de percentiles extremos (se tratarán como atípicos los percentiles extremos 0.5). (Calviño Martínez, 2020)

Si observamos la asimetría de las variables:

Variable	Mediana	Asimetría ▼
NumDealsPurchases	2	2.418569
MntSweetProducts	8	2.136081
MntFruits	8	2.102063
MntMeatProducts	67	2.083233
MntFishProducts	12	1.919769
MntGoldProds	24	1.886106
NumCatalogPurcha...	2	1.880989
NumWebPurchases	4	1.382794
MntWines	173	1.175771
Monetary	396	0.860841
NumStorePurchases	5	0.702237
REP_Income	51373	0.34735
Frequency	12	0.297114
NumWebVisitsMonth	6	0.207926
REP_Age	44	0.093266
Recency	49	-0.00199
Antiquity	540	-0.01522

Tabla 6 - Tabla para observar asimetría de las variables

Todas las variables a tratar son simétricas (el coeficiente de asimetría se encuentra entre -1 y 1). Por esta razón, los datos atípicos se tratarán con el método de desviación estándar mencionado anteriormente.

Se encontraron 7 observaciones atípicas en la variable *Income* y 5 observaciones atípicas en la variable *Monetary.* Con la utilización del nodo reemplazo, estas observaciones fueron remplazadas por valores ausentes que serán tratados en el próximo paso.

Esto se puede observar en la siguiente tabla:

Cuentas de reemplazo total

Variable	Entrenamiento ▼
REP_Income	7
Monetary	5
Antiquity	0
Frequency	0
Recency	0

Tabla 7 - Cantidad de observaciones con datos atípicos en las 5 variables que se van a utilizar

4.3.2.2. Tratamiento de datos faltantes:

En el caso que hubiera habido muchos datos faltantes, debería haberme fijado, con la creación de la variable numMissing, si algún cliente presentaba más de la mitad de las variables con datos faltantes. En este caso, únicamente dos variables presentan datos faltantes, por lo tanto, a lo sumo un cliente puede tener dos variables sin datos. Asimismo, si hubiese muchos datos faltantes en el dataset la variable numMissing podría estar dando información. Como no es el caso, esta variable no se va a crear.

Las variables de clase no tienen ningún dato ausente, por lo tanto, no es necesario imputarlas. En caso de que hubiera habido una variable de clase con entre un 5% y un 50% de datos faltantes, se hubiera creado a partir del nodo imputar o el nodo reemplazo

una nueva categoría válida para esos datos: “no consta” y en caso de que hubiera habido alguna variable con menos de un 5% de datos faltantes, se los hubiera imputado mediante una distribución aleatoria.

En cuanto a las variables de intervalo, no hay ninguna variable que tenga demasiados datos faltantes (+50%) como para que tenga que ser rechazada, ni ninguna variable con más de un 5% de datos faltantes como para tener que crearles variables de tipo único. La variable que presenta un mayor número de datos faltantes es la variable *Income* que presenta 32 datos faltantes. En el caso que hubiera alguna variable de intervalo con más de un 5% de datos faltantes (y menos de 50%), se hubieran creado con el nodo imputar, variables de tipo único en donde valen 1 en caso de que el dato haya sido inventado, y 0 en caso de que no. No es el caso, por lo tanto, no fueron creadas variables de tipo único.

Hay distintas formas para “rellenar los huecos” de los datos sin la necesidad de eliminar observaciones o variables para quitar los valores ausentes de la base de datos. Se puede hacer una imputación por modelos, o una imputación simple (media, mediana, distribución, etc.). Yo voy a utilizar una imputación simple. Voy a hacer que los datos faltantes, sean completados mediante una distribución aleatoria. De esta forma, el nodo tiene en cuenta la distribución de las variables y asigna valores a los datos faltantes de forma aleatoria. Esto me lleva a que no se desvirtúen los datos, manteniendo la forma.

Una vez imputados los datos faltantes con el nodo imputar, mediante la distribución aleatoria, la base de datos no presenta datos faltantes (configuración del nodo en el Anexo).

4.3.3. Análisis exploratorio de los datos ya depurados

Si se observa la matriz de correlaciones:

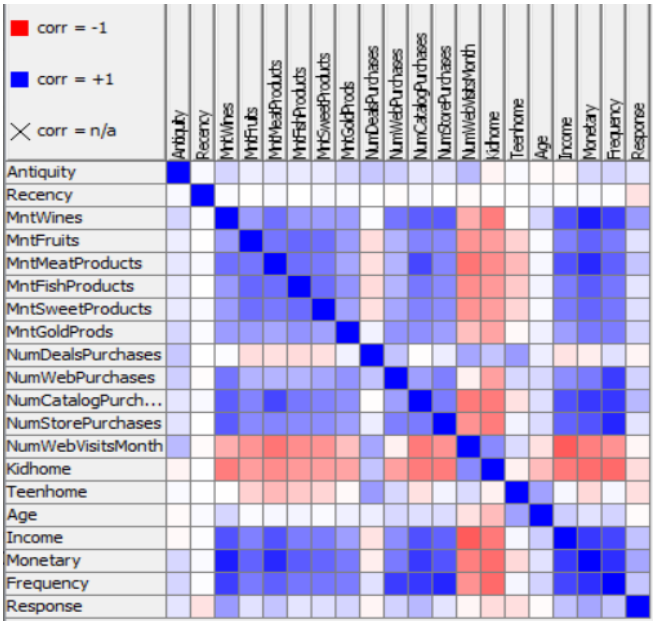


Ilustración 10 - Matriz de correlación entre variables

Como era de esperarse, la variable Monetary, está fuertemente correlacionada con la cantidad consumida de cada producto independientemente. Esto era de esperarse, ya que Monetary fue creada a partir de ellas.

Frequency y Monetary también están fuertemente relacionadas. Esto resulta intuitivo, ya que sería extraño que un cliente alcance un Monetary alto con una única compra. De la misma manera, también sería extraño que un cliente consuma frecuentemente y no logre alcanzar un Monetary alto.

Tampoco resulta extraño que Income esté relacionada con Frequency y Monetary. A mayor ingreso, mayor gasto y de manera más frecuente.

Por otra parte, para mi sorpresa, la variable *kidhome* está fuertemente relacionada pero negativamente con las variables *Monetary*, *Frequency*, *Income* y *NumWebVisitsMonth*. Esto quiere decir que en el caso que haya niños en el hogar, el ingreso, el gasto, la frecuencia en la que consumen y la cantidad de veces que se visita la página web será menor que si no hay niños en el hogar. No ocurre lo mismo con el consumo en tiendas físicas. Si hay niños en el hogar, la relación con el consumo en tienda físicas es positiva. En este sentido, el consumo en tiendas físicas es mayor si hay niños en el hogar que si no los hay. La razón por la cual sucede esto no es para nada intuitiva.

Asimismo, cabe destacar que la recencia no presenta ningún tipo de relación con ninguna variable dentro de la base de datos.

4.3.4. Normalización de variables

Cuando las variables que se van a utilizar para construir los modelos de Machine Learning (tanto supervisados como no supervisados) son numéricas, la escala en la que se miden y la magnitud de la varianza pueden influir en el modelo. En este sentido, se debe tratar de igualar de alguna forma las variables, que todas las variables se encuentren en un rango definido. (Rodrigo, Ciencia de Datos, 2020)

El proceso de transformación de escala en la distribución de una variable es conocido como normalización o estandarización. El objetivo es poder hacer comparaciones entre variables eliminando efectos de influencias. (Rodó, 2019).

Hay dos estrategias ampliamente conocidas para normalizar variables:

- Normalización Z-score: consiste en dividir las variables entre su desviación típica después de restarles su media. De esta forma, los datos pasan a tener una distribución normal.

Ecuación 4 - Normalización Z-score

$$z = \frac{x - \mu}{\sigma}$$

- Estandarización min-max: consiste en transformar los datos de forma que estén dentro del rango [0, 1]. Cada entrada se normaliza entre unos límites definidos.

Ecuación 5 - Normalización Min - Max

$$Xnorm = \frac{X - Xmin}{Xmaz - Xmin}$$

Cada una de las estrategias tienen sus cosas buenas y sus cosas malas.

En este caso, se normalizaron las variables que van a ser utilizadas para generar los clústeres con la estrategia Min-Max. Para dicho procedimiento, se utilizó el nodo

Normalizer de KNIME. Se normalizaron las variables Age, Income, Antiquity, Recency, Frequency y Monetary.

Asimismo, cabe destacar que, para evitar confusiones con la recencia, una vez normalizada, se invirtió la variable. Se creó una "nueva recencia" a partir de la anterior, (1 – recencia), de forma que cuanto más cercano al 1, mejor será el cliente (compró más recientemente), y cuanto más cercano a cero sea el valor que tome la recencia, más antigua será su última compra.

4.4. Modelado y Evaluación

4.4.1. Modelo RFM

Se va a comenzar analizando el modelo RFM. En primer lugar, se calculará el valor del cliente en KNIME. Luego, se llevarán los datos a R para encontrar los clústeres óptimos.

4.4.1.1. RFM score en KNIME:

Se van a dividir cada una de las 3 dimensiones del modelo RFM en 5 grupos (quintiles). Cada uno de los 5 grupos tendrán la misma cantidad de observaciones. El 20% de los clientes con Frecuencia más alta tendrán puntaje 5, el 20% siguiente 4 y así sucesivamente. Lo mismo pasará con el valor monetario. En cuanto a la recencia, el 20% de los clientes con Recencia más baja tendrán puntaje 5, el 20% siguiente 4 y así sucesivamente. Una vez calculado el puntaje en cada una de las dimensiones individualmente, se calculará el puntaje RFM total siguiendo la siguiente fórmula:

Ecuación 6 - RFM Score

$$RFM\ score = Puntaje\ Recencia * \frac{1}{3} + Puntaje\ Frecuencia * \frac{1}{3} + Puntaje\ Valor\ Monetario * \frac{1}{3}$$

En la tabla a continuación, se muestran el valor mínimo, máximo y la media de cada quintil para cada una de las 3 dimensiones, *Recencia, Frecuencia y Valor monetario:*

Puntaje RFM	Recencia (días)			Frecuencia (compras)			Valor Monetario (dólares)		
	Min	Max	Media	Min	Max	Media	Min	Max	Media
1 (Muy bajo)	80	99	89	0	5	4	5	55	35
2 (Bajo)	60	79	70	6	10	7	56	198	92
3 (Medio)	40	59	50	11	15	13	199	637	399
4 (Alto)	20	39	29	16	20	18	639	1174	928
5 (Muy alto)	0	19	9	21	32	23	1175	2352	1536

Tabla 8 - Escalas definidas para cada una de las 3 dimensiones, Recencia, Frecuencia y Valor monetario

En base a la tabla anterior se puede decir lo siguiente:

Los clientes con un puntaje 5 en recencia, a lo sumo hace 19 días que consumieron en la empresa. En promedio, los clientes con puntaje 5 tienen una recencia de 9 días. Por otra parte, los clientes con puntaje 1 llevan entre 80 y 99 días sin consumir en la empresa. El promedio de los clientes del peor quintil es de 89 días sin consumir (aproximadamente 3 meses).

Los clientes con puntaje 5 en frecuencia consumieron entre 21 y 32 veces en la empresa. En promedio, los clientes de este quintil consumieron 23 veces. No resulta extraño que el promedio se encuentre más cerca del mínimo que del máximo, ya que son pocos los clientes que consumieron demasiadas veces. Por otra parte, si se observa

el peor quintil de la frecuencia, hay clientes en la empresa que no consumieron ni una vez. Esto podría ser un error. A pesar de ello, los clientes del peor quintil consumieron en promedio 4 veces.

Por último, los clientes con mayor puntaje en valor monetario gastaron entre 1175 y 2352 dólares en la empresa. El promedio de los clientes con puntaje 5 en valor monetario es de 1536 dólares. Al igual que con la frecuencia, hay clientes que no gastaron nada en la empresa y se encuentran en el peor quintil. También podría ser un error. Los clientes que se encuentran en el peor quintil del valor monetario gastaron como máximo 55 dólares. En promedio, los clientes de ese quintil gastaron 35 dólares.

Asimismo, una vez calculado el RFM score, se puede observar en la tabla a continuación la cantidad de clientes con cada uno de los puntajes.

Row ID	count
1.0	81
1.3333333333333333	106
1.6666666666666665	184
1.9999999999999998	107
2.0	94
2.333333333333333	145
2.3333333333333335	101
2.6666666666666665	163
2.666666666666667	32
2.9999999999999996	17
3.0	217

Row ID	count
3.0000000000000004	2
3.333333333333333	189
3.3333333333333335	28
3.666666666666666	60
3.6666666666666665	14
3.666666666666667	195
4.0	189
4.333333333333333	101
4.333333333333334	60
4.666666666666667	130
5.0	25

Tabla 9 - Cantidad de clientes cada uno de los puntajes globales (RFM score)

Se puede ver que hay más clientes con puntaje 1 que clientes con puntaje 5. Asimismo, hay 1247 clientes con puntajes entre 1 y 3 incluido y 993 clientes con puntajes mayores a 3. En este sentido, hay más clientes malos que buenos. Se puede observar lo mismo en el siguiente gráfico de barras:

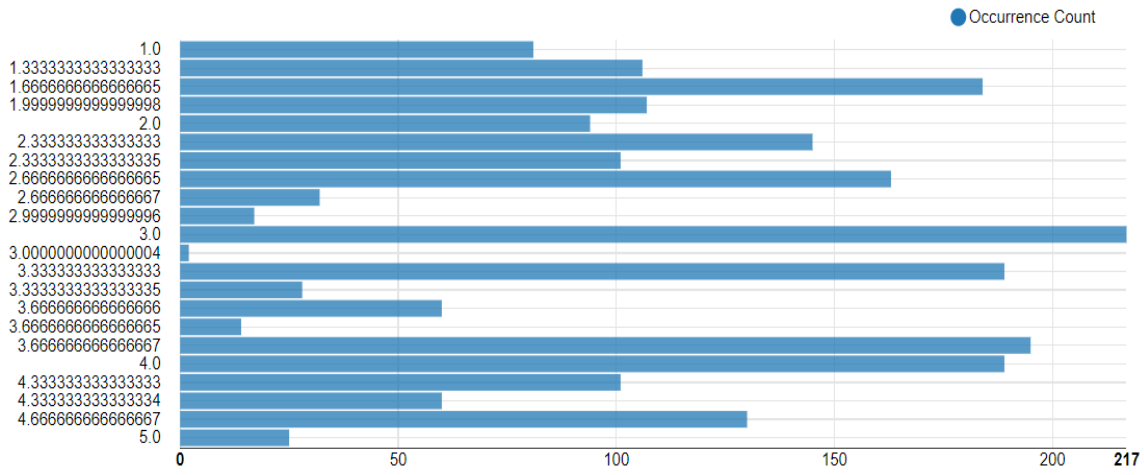


Ilustración 11 - Gráfica de barras, cantidad de clientes cada uno de los puntajes globales (RFM score)

De todas maneras, como fue mencionado en el capítulo anterior, para sacar el máximo provecho posible del modelo RFM, se van a realizar clústeres de clientes para segmentarlos en base a las 3 dimensiones del modelo.

4.4.1.2. Segmentación de clientes en base a las tres dimensiones del modelo RFM en R:

La base de datos fue exportada de KNIME con las variables ya normalizadas e importada en R para la creación de los clústeres.

Si se observa gráficamente la distribución de las variables:

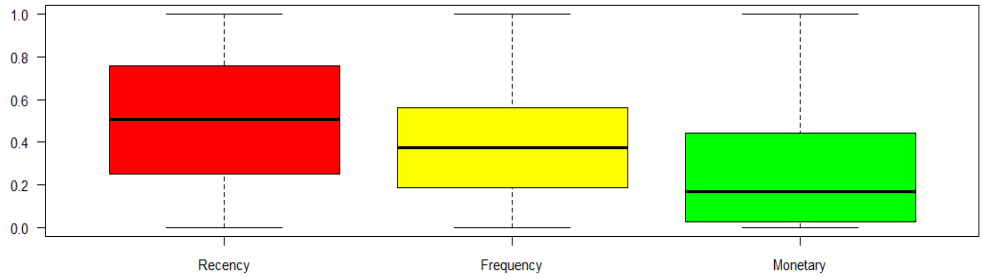


Ilustración 12 - Gráfico de caja de las dimensiones del modelo RFM

La recencia es la variable con valores más uniformes. Si se observa el valor monetario, hay más observaciones que toman valores chicos. Esto no es llamativo, ya que hay pocos clientes que consumen mucho. De todas maneras, los datos atípicos fueron tratados en la etapa de depuración de datos.

La herramienta R tiene paquetes que resultan de mucha utilidad a la hora de generar clústeres de clientes. La librería "*stats*" tiene la función *kmeans ()* para la creación de clústeres con el algoritmo k-media.

De todas maneras, antes de comenzar a analizar los clústeres con k-media, me gustaría evaluar si hay indicios de que realmente existe algún tipo de agrupación en ellos. La razón por la cual es necesario hacer esto, es porque los algoritmos de clustering imponen una clasificación, aunque no existan clústeres relevantes en los datos. Hay varias formas de realizar esto. Yo voy a utilizar el **estadístico Hopkins**.

En R existe la función *get_clust_tendency ()* del paquete *factoextra* para calcular el valor del estadístico.

Se calculó el valor del estadístico únicamente para las variables Recencia, Frecuencia y Valor Monetario. El estadístico toma un valor de **0.8279156**. Esto significa que la sumatoria de las distancias entre los puntos de mi base de datos y sus vecinos más cercanos son mucho menores a la sumatoria de las distancias de la base de datos aleatoria con sus vecinos más cercanos. Por lo tanto, hay indicios que se podrían formar clústeres buenos con esas 3 variables.

Una vez que confirmamos que los datos no se distribuyen de forma uniforme y, por lo tanto, se pueden aplicar algoritmos de clustering, se va a observar qué distribución tienen. Para ello, se va a realizar un gráfico de dispersión con la función *fviz_pca_ind ()* del paquete *"factoextra".* Previo a ello, es necesario reducir la dimensionalidad (tengo tres dimensiones y necesito tener únicamente dos). Esto se va a hacer aplicando análisis de componentes principales (PCA) con la función de R *prcomp ().*

El gráfico resultante es el siguiente:

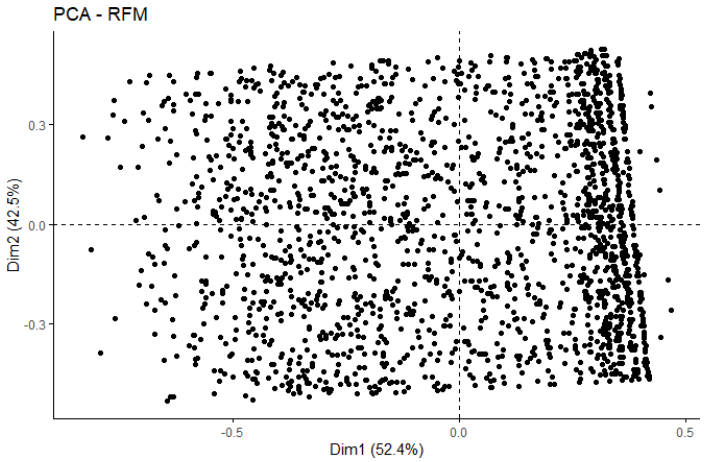


Ilustración 13 - Gráfico de dispersión RFM (PCA)

En el eje horizontal se muestra la dimensión 1 con un 52,4% de la variabilidad y en el eje vertical se muestra la dimensión 2 que representa un 42,5% de variabilidad. En este sentido la reducción en dos componentes está consiguiendo representar casi el 95% de la información.

Se prueba que tal funciona el algoritmo k-medias para k=4 y se vuelve a representar el gráfico anterior:

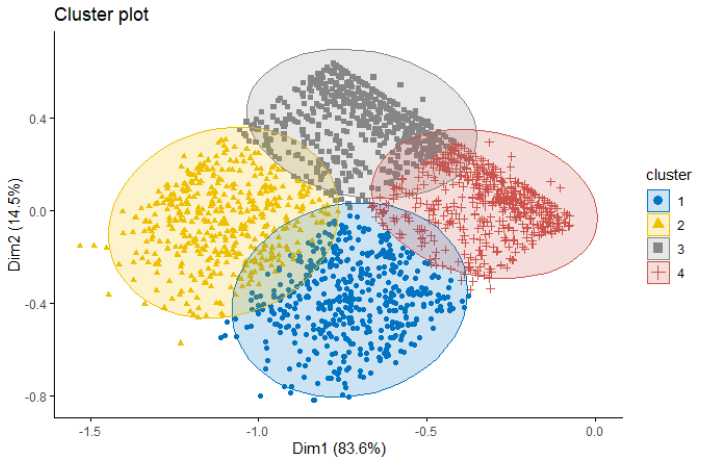


Ilustración 14 - Gráfico de dispersión RFM (PCA) con visualización de 4 clústeres realizados con k-media

Se puede observar que el algoritmo k-media funcionaría bien en mis datos.

Se puede observar que a pesar de que los clústeres formados no sean esféricos, se podría decir que tienen una forma separable y definida y por lo tanto que el algoritmo k-medias funcionaría bien en los datos.

Por otra parte, previo a la creación de los clústeres, se va a intentar buscar la cantidad de clústeres optima. La librería *"factoextra"* brinda funciones fáciles de utilizar para realizar las gráficas de Elbow y Silhouette. La idea de utilizar esta función para encontrar el número de clúster óptimo, es no tener que ir probando uno a uno cual sería el mejor número de clústeres.

Si se observa el gráfico Elbow:

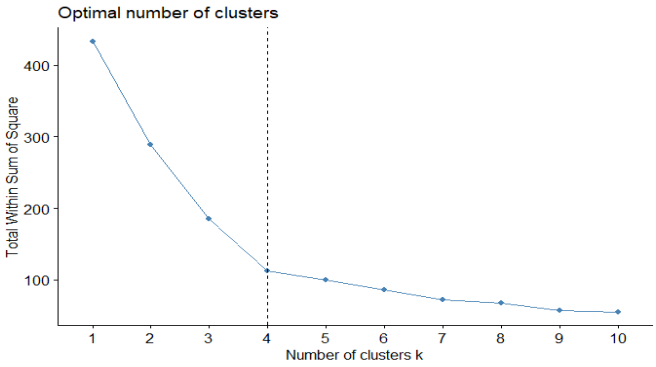


Ilustración 15 - Gráfica Elbow RFM en R

Si se observa el gráfico de Average Silhouette:

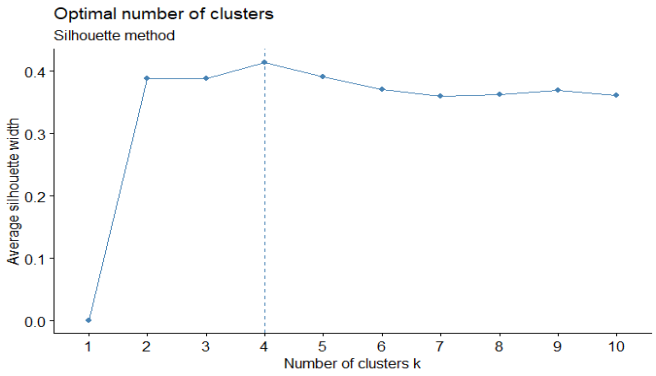


Ilustración 16 - Gráfica Average Silhouette RFM en R

Ambos métodos arrojan el mismo resultado considerando que el numero óptimo de clústeres es 4. Para k=4 la silueta media alcanza un valor de 0.41, lo cual es bueno.

Asimismo, en el paquete "*NbClust*" de R, existe una función, también muy fácil de utilizar, en donde evalúa el número de clústeres óptimo por 26 métodos distintos. Una vez evaluados los 26 métodos, la función sugiere como cantidad óptima de clústeres, aquella que se repitió más veces entre los 26 métodos utilizados. La función brinda un gráfico de barras donde da el número de clústeres por la frecuencia. Siendo la frecuencia la cantidad de veces que ese número de clústeres fue elegido:

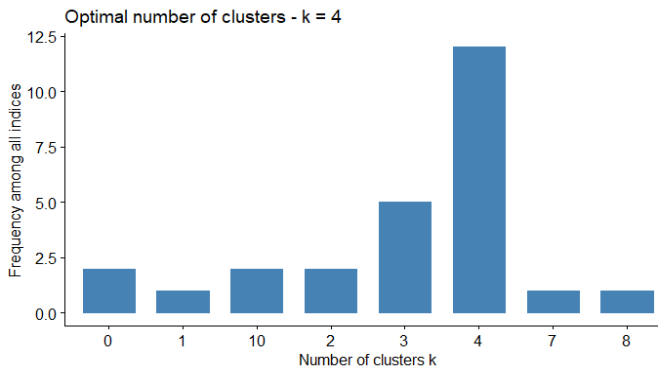


Ilustración 17 - Gráfico de frecuencia de la cantidad de clústeres en base a la cantidad de veces que fueron elegidos

Se puede saber que de los 26 métodos utilizados para encontrar la cantidad de clústeres óptima:

- 12 propusieron 4 como el mejor número de clústeres.
- 5 propusieron 3 como el mejor número de clústeres.
- 2 propusieron 0, 2 y 10 como el mejor número de clústeres.
- 1 propuso 1, 7 y 8 como el mejor número de clústeres.

Se podría concluir entonces que el número óptimo de clústeres es 4. De todas maneras, más allá de que 4 sea el número de clústeres que minimice la varianza interna y maximiza la silueta promedio, es probable que, si se aumenta el número de clústeres, se puedan conseguir segmentos interesantes con sentido de negocio. A pesar de que tal vez los clústeres formados presenten un poco más de ruido, tal vez es de más utilidad crear más segmentos. Se va a probar segmentar a los clientes formando 4, 5, 6 y 7 clústeres para observar que segmentos se forman.

Para k=4:

```
Cluster means:
    Recency Frequency   Monetary
1 0.7683734 0.6182651 0.50096051
2 0.2460435 0.2233306 0.06263801
3 0.2630163 0.6023166 0.50132593
4 0.7531056 0.2198661 0.06180151
```

Cantidad Obs.	Perfil	Silueta
464	Vip	0,33
614	Peores	0,48
518	Churn	0,33
644	Nuevos	0,48

Tabla 10 - Centroide de los clústeres formados en R para k=4

Se podría decir que las observaciones están bien distribuidas entre los clústeres. No hay ningún clúster con demasiadas observaciones, ni ningún clúster con muy pocas observaciones. Asimismo, no hay grandes diferencias entre las siluetas de los clústeres.

Clúster 1: corresponde a los clientes **VIP**, con recencia, frecuencia y valor monetario alto. Son los mejores clientes de la empresa.

Clúster 2: corresponde a los **peores** clientes de la empresa. Recencia, frecuencia y valor monetario bajo.

Clúster 3: corresponde a los **Churn**. Son clientes que solían ser buenos, pero por una cosa u otra hace mucho que no consumen en la empresa

Clúster 4: Corresponde a los clientes **nuevos** o clientes que se están reactivando. Tiene recencia alta, pero valor monetario y frecuencia relativamente baja. Para corroborar si efectivamente son nuevos, se debería observar cuáles de estos clientes tienen antigüedad baja.

Si en lugar de formar 4 clústeres como fue recomendado, decidiera realizar 5, para ver si se genera algún segmento interesante, tendríamos lo siguiente:

Para K=5:

```
Cluster means:
    Recency Frequency   Monetary
1 0.5042660 0.2188258 0.06068065
2 0.1617370 0.2255532 0.06402452
3 0.2611833 0.6034736 0.50494657
4 0.8387097 0.2245104 0.06390254
5 0.7656131 0.6179435 0.50079213
```

Cantidad Obs.	Perfil	Silueta
412	Ocacionales	0,35
418	Peores	0,47
511	Churn	0,32
434	Nuevos	0,43
465	Vip	0,32

Tabla 11 - Centroide de los clústeres formados en R para k=5

En este caso se obtienen los 4 clústeres similares a los anteriores (cuando k=4), pero se adiciona un nuevo clúster. El clúster de los clientes ocasionales. La silueta promedio de este modelo es de 0.38, algo inferior que para k=4 pero sigue siendo buena.

La forma en la que fueron formados estos clústeres no parece tan acertada desde el punto de vista de negocio. Los clústeres 1 (Ocasionales), 2 (Peores) y 4 (Nuevos) se solapan un poco. Se podría decir que el clúster 1 es una mezcla entre el clúster 2 y el clúster 4, ya que fueron etiquetados como “ocasionales” pero podrían ser “nuevos” o de los “peores”. De todas maneras, las observaciones están bien distribuidas entre los clústeres. No hay ningún clúster con muchas observaciones ni ningún clúster con pocas observaciones.

Para K=6:

```
Cluster means:
    Recency Frequency   Monetary
1 0.7866487 0.6143882 0.49984085
2 0.1582231 0.2076066 0.05221356
3 0.3202897 0.5318627 0.32545033
4 0.5150172 0.1864195 0.04192201
5 0.2536022 0.6575138 0.65149972
6 0.8426066 0.2282512 0.06555192
```

Cantidad Obs.	Perfil	Silueta
427	Vip	0,32
387	Peores	0,5
357	Leales	0,25
376	Ocacionales	0,41
272	Churn	0,27
421	Nuevos	0,42

Tabla 12 - Centroide de los clústeres formados en R para k=6

En primer lugar, se puede decir que las observaciones están bien distribuidas entre los clústeres. La silueta media de este modelo es de 0.37, apenas por debajo que cuando k=5.

La gran diferencia con el modelo formado con k=5 es que, en este caso, se adiciona un clúster que podría resultar interesante. El clúster 3, los clientes leales. Son clientes que no son tan buenos como los VIP, o como supieron ser los Churn, pero que podrían ser potenciales clientes, en crecimiento.

De todas maneras, al observar las siluetas, se puede ver que el clúster 2 es el clúster con observaciones más homogéneas y al mismo tiempo es el clúster de menos utilidad si se piensa desde el sentido de negocio. Es el clúster de los peores clientes, con frecuencia, recencia y valor monetario bajo. A pesar de que resulte importante identificar cuáles son los peores clientes de la empresa, es probable que las personas que se encuentren en ese clúster no puedan ser recuperadas. Invertir dinero en tratar de conseguirlos sería el último en la lista de prioridades.

Para K=7:

```
Cluster means:
    Recency Frequency   Monetary
1 0.3100952 0.5373013 0.33155596
2 0.5041734 0.1857493 0.04030654
3 0.1575126 0.2063802 0.05156738
4 0.8409301 0.1844702 0.04051497
5 0.7548015 0.6465669 0.61581462
6 0.2293333 0.6502500 0.64754666
7 0.7976666 0.5168362 0.27162335
```

Cantidad Obs.	Perfil	Silueta
346	Leal	0,23
357	Ocacionales	0,42
384	Peores	0,49
361	Nuevos	0,49
284	Vip	0,24
250	Churn	0,28
258	Crecimiento	0,3

Tabla 13 - Centroide de los clústeres formados en R para k=7

Cuando k=7 se adiciona a los clústeres anteriores un nuevo segmento que denominé Crecimiento. Estos clientes podrían aumentar su consumo con campañas cross selling y up selling. Tienen una recencia alta, una frecuencia media/alta, pero valor monetario

medio/bajo. Desde el punto de vista de negocio, es un segmento interesante para aplicarle campañas de marketing.

En este modelo las observaciones también están bien distribuidas entre los clústeres y la silueta media es de 0.36.

En este caso, 3 de los 7 clústeres formados parecen ser buenos, pero 4 de ellos no son tan buenos. De los tres clústeres buenos (2 (ocasionales), 3 (peores) y 4 (nuevos)), el 2 y el 4 podrían ser de utilidad desde el punto de vista del negocio. Los clientes en el clúster 7 son los que solían ser muy buenos clientes, pero por alguna cosa u otra dejaron de consumir en la empresa. Intentar recuperar a estos clientes serían una de las primeras cosas que la empresa debería poner en marcha. De todas maneras, el nuevo segmento encontrado (clúster 7, en crecimiento), podría ser muy útil desde el punto de vista de negocio.

Por lo tanto, a pesar de que la silueta media sea bastante peor a la obtenida con k=4, no descartaría la posibilidad de segmentar a los clientes en estos 7 clústeres. Ambos modelos serán evaluados en mayor profundidad en el apartado de "evaluación" para poder decidir con cuál de los dos quedarnos.

4.4.2. Evaluación modelo RFM

Del apartado de modelado se pudo concluir que k=4 era el número óptimo de clústeres desde el punto de vista de la distribución de los datos, sin tener en cuenta el sentido de negocio. Para **k=4** se maximizaba la silueta media y si se observaba la silueta de cada uno de ellos era similar en los 4 (ningún clúster tenía una silueta mucho más alta que otro). De todas maneras, también se pudo observar que para **k=7** se generaban nuevos segmentos que podían resultar interesantes desde el punto de vista de negocio. Por esta razón, se va a comparar en más profundidad ambos modelos para poder elegir cuál de las dos segmentaciones utilizar.

Si se observan gráficamente los clústeres formados para **k=4**:

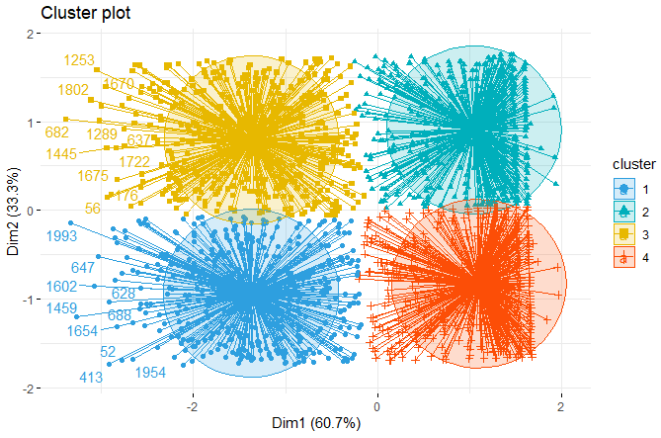


Ilustración 18 - Visualización de los clústeres RFM para k=4

A simple vista, parecería ser que los clústeres quedan relativamente bien formados. Asimismo, si se observa gráficamente la silueta de cada uno de los clústeres:

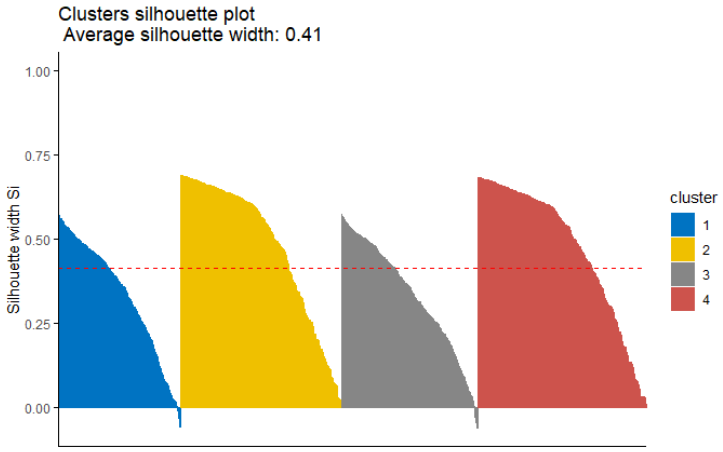


Ilustración 19 - Gráfica de la silueta de los clústeres formados (RFM k=4)

La silueta media del modelo es de 0.41 lo cual es muy bueno, ya que, a pesar de que la silueta media puede tomar valores entre -1 a 1 y cuánto más cerca de 1 se encuentre, mejor es, no es muy normal en la práctica encontrarse con valores demasiados altos de silueta media. En este sentido, un valor de 0.41 es razonablemente bueno. Se puede ver también que todos ellos presentan silueta similar. No hay un clúster que esté sustancialmente mejor formado que otro. En este sentido, las observaciones están cerca de su propio clúster y lejos de otros.

Por otra parte, se va a observar qué tan amplio es el rango de valores que toma cada variable en cada uno de los 4 clústeres formados. A pesar de que conocemos el centroide (que es equivalente a la media), no se conoce exactamente la amplitud de valores que toman.

Cluster RFM	Recencia			Frecuencia			Valor Monetario		
	Min	Max	Media	Min	Max	Media	Min	Max	Media
1 (Vip)	0	48	22	10	32	20	277	2352	1179
2 (Peores)	50	99	74	0	16	7	5	801	146
3 (Churn)	48	99	73	1	32	19	55	2346	1179
4 (Nuevos)	0	49	24	0	17	7	6	825	146

Tabla 14 - Tabla con el mínimo, máximo y media de cada dimensión del modelo RFM en k=4

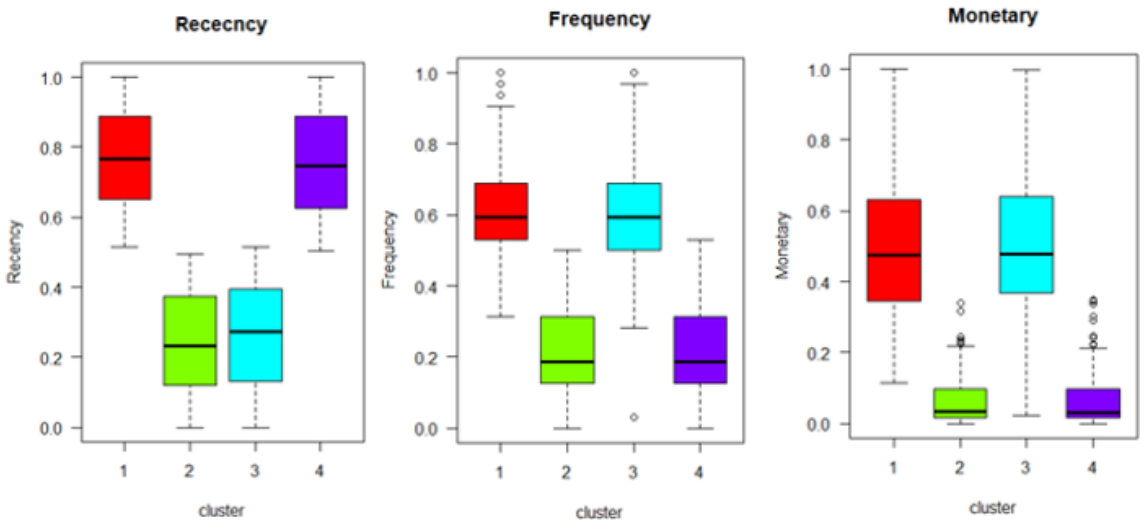


Ilustración 20 - Gráfico de cajas de las variables Monetary, Frequency y Recency para cada clúster

En primer lugar, me gustaría notar que en la variable Monetary el clúster 1 (VIP) y el clúster 3 (Churn) hay clientes con valor monetario muy pequeño y clientes con valor monetario muy alto. La amplitud de valores que toman es muy grande. Esto puede ser un indicio de que los clústeres no tienen una forma perfecta. De todas maneras, resulta intuitivo que esto suceda, ya que hay muchos más clientes con Monetary bajo qué alto. Asimismo, si se observa la ilustración 19, se puede ver que justamente los clústeres 1 y 3 tienen observaciones con valor de silueta negativa. Este es un indicativo que estas observaciones podrían estar mal asignadas dentro del clúster y pueden ser las que están causando la amplitud tan grande en el valor monetario. Podría resultar interesante estudiar porque esos clientes han sido clasificados en esos segmentos que, en términos de negocios, son tan importantes.

Si se observan las similitudes y diferencias existentes entre los clústeres, se podría decir que los clientes pertenecientes al clúster 1 (Vip) y 3 (Churn) tienen valor monetario medio / alto y frecuencia alta. La diferencia entre ellos es que mientras los clientes en el clúster 1 tienen recencia alta, los clientes en el clúster 3 tienen recencia baja.

Por otra parte, sucede lo mismo con los clústeres 2 (peores) y 4 (nuevos). Ambos tienen valor monetario y frecuencia baja. La diferencia entre ellos es que mientras los clientes nuevos tienen recencia alta, los peores clientes tienen recencia baja.

Se podría decir que los clústeres están bien definidos. De todas maneras, se van a tomar al azar 8 clientes de cada uno de los clústeres para ver de forma más clara si son homogéneos entre sí, y si los clústeres formados son heterogéneos:

Recency	Frequency	Monetary	cluster	Recency	Frequency	Monetary	cluster
38	4	27	4	81	16	507	3
26	6	53	4	53	16	1,105	3
32	8	169	4	78	14	974	3
19	5	46	4	83	12	892	3
11	3	19	4	64	16	562	3
38	4	46	4	69	19	1,334	3
20	11	317	4	59	21	1,141	3
41	11	316	4	95	20	1,538	3
4	10	257	4	92	24	1,152	3

Recency	Frequency	Monetary	cluster	Recency	Frequency	Monetary	cluster
53	6	78	2	23	18	639	1
64	10	244	2	25	23	1,665	1
52	14	425	2	44	12	1,575	1
56	4	34	2	0	21	1,208	1
94	4	29	2	10	23	1,034	1
56	14	396	2	29	23	1,088	1
86	6	55	2	23	18	653	1
57	4	32	2	11	17	602	1
77	5	67	2	48	29	1,366	1
62	15	496	2	24	24	1,001	1

Tabla 15 - Observaciones pertenecientes a cada clúster con k=4

En el clúster 4, que corresponde a los clientes nuevos, se puede observar que los clientes toman valores similares. Algunos tienen una frecuencia y valor monetario algo mayor que los otros, pero se podría decir que no hay grandes diferencias entre clientes. Si se los compara con los clientes del clúster 2 se puede ver como claramente la diferencia radica en la recencia. Mientras que en el clúster 4 todos los clientes tienen una recencia menor de 40, los clientes del clúster 2 tienen una recencia superior a 50. Lo mismo sucede con los clústeres 1 y 3, tienen frecuencia y valor monetario similar, pero la recencia del clúster 1 es menor que la del clúster 3.

Si ahora se pasa a examinar en mayor profundidad **k=7**:

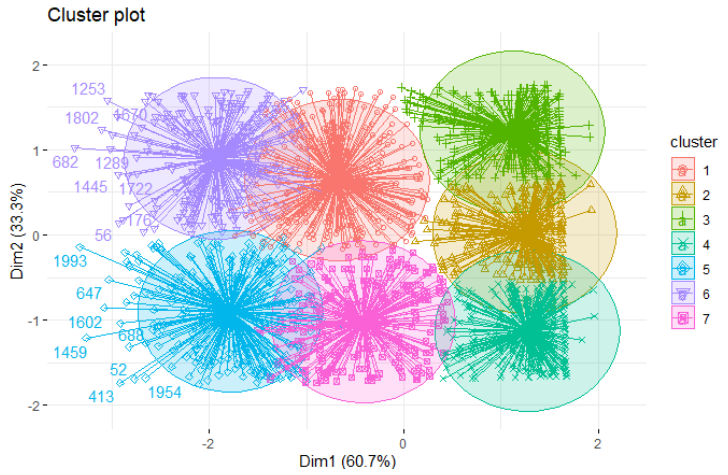


Ilustración 21 - Visualización de los clústeres RFM para k=7

A simple vista, parecería ser que los clústeres están correctos.

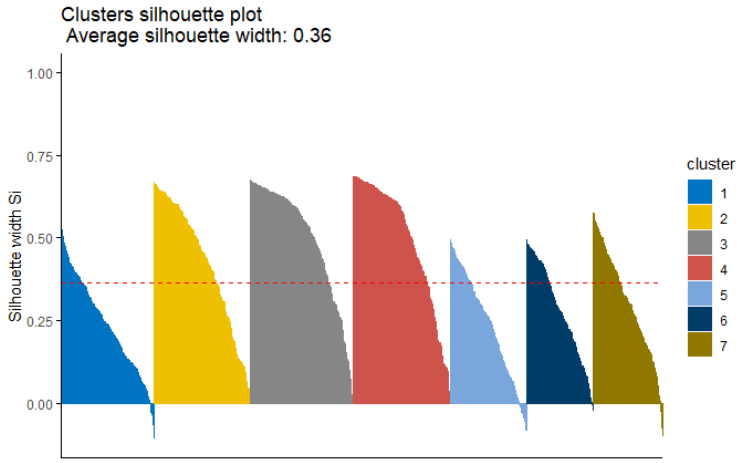


Ilustración 22 - Gráfica de la silueta de los clústeres formados (RFM k=7)

Como ya fue mencionado en la etapa de modelado, el modelo alcanza una silueta media de 0.36 lo cual, a pesar de que es peor que la silueta para k=4, sigue resultando relativamente buena. Los clústeres 2 (Ocasionales), 3 (Peores) y 4 (Nuevos), son sin duda los clústeres con mayor silueta.

Al igual que para k=4 se va a observar que tan amplio es el rango de valores que toma cada variable en cada uno de los 7 clústeres formados:

Cluster RFM	Recencia			Frecuencia			Valor Monetario		
	Min	Max	Media	Min	Max	Media	Min	Max	Media
1 (Leal)	44	99	68	1	28	17	55	679	787
2 (Ocacional)	33	66	49	0	14	6	5	393	100
3 (Peores)	66	99	84	0	14	7	6	801	125
4 (Nuevos)	0	32	16	0	12	6	6	373	100
5 (Vip)	0	49	24	10	32	21	863	2352	1438
6 (Churn)	51	99	77	10	32	21	1003	2346	1508
7 (Crecimiento)	0	44	20	9	27	16	277	1190	639

Tabla 16 - Tabla con el mínimo, máximo y media de cada dimensión del modelo RFM en k=7

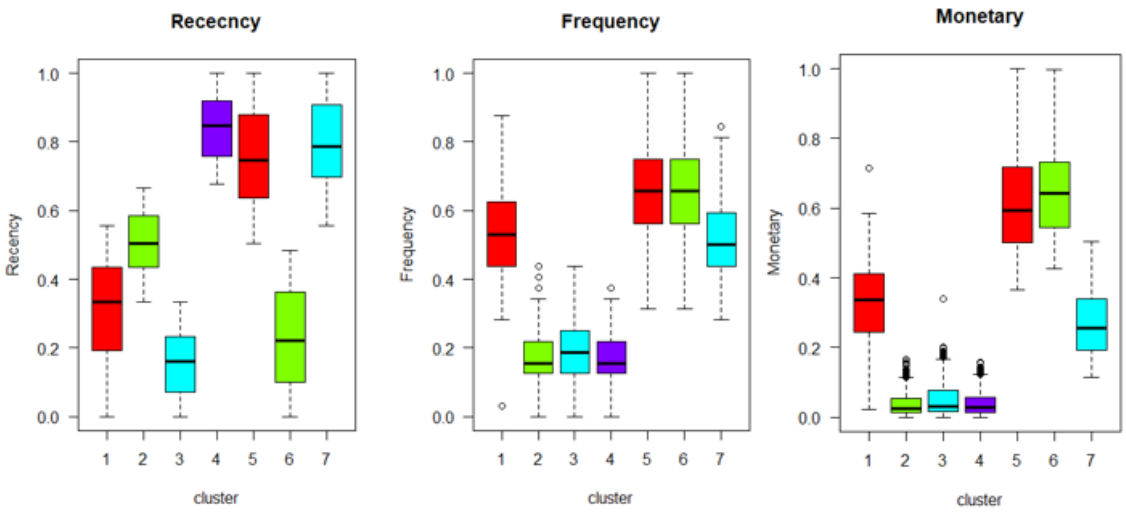


Ilustración 23 - Gráfico de cajas de las variables Monetary, Frequency y Recency para cada clúster (k=7)

En este caso es interesante notar las diferencias existentes entre clústeres. Los clientes pertenecientes a los clústeres 2 (Ocasional), 3 (Peores) y 4 (Nuevos) tienen una frecuencia y un valor monetario muy similar, sin embargo, la diferencia entre ellos radica en la Recencia. Los clientes Nuevos, tienen una recencia alta, los peores tienen una recencia baja y los ocasionales una recencia media. En este caso, tal vez se podría aplicar algún tipo de campaña para que los clientes pertenecientes al clúster 2 se pasen al clúster 4 (como clientes que se están reactivando) y no al clúster 3. Una campaña de retención.

Por otra parte, se encuentran los clientes pertenecientes a los clústeres 5 (Vip) y 6 (Churn). Ambos tienen frecuencia y valor monetario alto. Al igual que en el caso anterior, la diferencia también radica en la recencia. Los clientes pertenecientes al clúster 6 tienen una recencia baja, mientras los pertenecientes al clúster 5 tienen una recencia alta. En este caso, es de vital importancia intentar reactivar a los clientes pertenecientes al clúster 6 para que vuelvan a ser parte del clúster 5.

Por último, se encuentran los clientes pertenecientes a los clústeres 1 (Leal) y 7 (crecimiento). Ambos presentan frecuencia y valor monetario medio. Una vez más, la diferencia entre ambos clústeres radica en la recencia. Mientras los clientes en crecimiento tienen recencia alta, los clientes leales tienen una recencia más bien media/baja. La razón por la cual fueron denominados así es porque los clientes en crecimiento están activos. Consumieron recientemente, lo que lleva a que sean más propensos a responder a ofertas de productos cross selling y up selling. Por otra parte, los clientes leales parece ser el tipo de cliente que consume en la empresa cuando lo necesita. Tal vez se podría decir que son ocasionales frecuentes. Siendo que no consumieron en la empresa recientemente, sería difícil identificar campaña cross selling y up selling.

En este caso, se van a tomar al azar 8 clientes de cada uno de los clústeres para ver de forma más clara y son homogéneos entre sí, y si los clústeres formados son heterogéneos:

Recency	Frequency	Monetary	cluster7
26	20	776	7
16	20	716	7
34	17	590	7
0	17	544	7
19	27	637	7
29	16	507	7
18	18	612	7
26	20	884	7

Recency	Frequency	Monetary	cluster7
91	22	2,279	6
98	16	1,779	6
85	26	1,804	6
78	17	1,862	6
73	18	1,383	6
56	18	1,286	6
95	22	1,192	6
90	26	1,229	6

Recency	Frequency	Monetary	cluster7
8	16	1,869	5
6	28	1,364	5
28	20	1,525	5
45	13	1,612	5
17	23	1,071	5
8	27	1,167	5
43	20	1,478	5
15	17	1,808	5

Recency	Frequency	Monetary	cluster7
8	12	318	4
20	10	253	4
11	6	76	4
21	5	62	4
10	3	9	4
25	5	68	4
30	7	173	4
30	7	99	4

Recency	Frequency	Monetary	cluster7
91	6	122	3
98	11	270	3
91	3	18	3
78	6	46	3
92	5	47	3
95	3	15	3
77	3	23	3
91	7	133	3

Recency	Frequency	Monetary	cluster7
48	6	49	2
39	6	75	2
64	10	215	2
54	4	47	2
57	6	143	2
64	5	55	2
41	5	81	2
47	14	382	2

Recency	Frequency	Monetary	cluster7
58	16	729	1
98	20	767	1
81	17	573	1
66	15	458	1
87	13	1,027	1
72	21	860	1
67	16	570	1
54	20	769	1

Tabla 17 - Observaciones pertenecientes a cada clúster con k=7

Se puede ver como efectivamente los clientes pertenecientes al clúster 7 (crecimiento) tienen un valor monetario medio y una frecuencia media / alta si se los compara con los demás clústeres, y una recencia relativamente más baja. Por otra parte, sin duda los clientes con valor monetario y frecuencia más alta son los pertenecientes a los clústeres 6 (Churn) y 5 (Vip) y la recencia es notablemente inferior en los clientes del clúster 5 (Vip) que en los clientes del clúster 6 (Churn). Asimismo, los clientes con valor monetario más bajo sin duda se encuentran en los clústeres 2 (Ocasionales), 3 (Peores) y 4 (Nuevos) y se puede notar la diferencia en la recencia de los clientes entre los 3 clústeres. Mientras los clientes nuevos a lo sumo consumieron hace 30 días, los clientes ocasionales entre 39 y 64 días y los peores clientes hace más de 78 días que no consumen. (Quiero aclarar que estos datos son una muestra y la misma fue elegida al azar). Por último, se encuentran los clientes del clúster 1. De la muestra, se puede ver que efectivamente son clientes con valor monetario medio y frecuencia media alta si se los compara con los clientes pertenecientes a los demás clústeres, y con una recencia más bien alta.

En este sentido, podría decir que los clústeres formados con k=7 parecerían ser correctos.

Para poder tomar la decisión sobre que segmentación elegir, me gustaría tener en cuenta lo siguiente: una buena segmentación debe proporcionar una serie de grupos de clientes que tenga las siguientes características:

- Cuantificables o medibles. Cada grupo que se obtengan ha de ser cuantificable: número de clientes que lo integra, descriptivos de las variables que lo define.
- Accesible. De tal manera que de una forma eficiente se puede llegar a dichos grupos tanto en el marketing directo que se haga, como en promoción más masiva mediante publicidad, etc. que permite seguir capturando leads de dicho segmento.

- Volumen mínimo. Aunque puedan existir segmentos con reducido número de clientes (por ejemplo, de altos ingresos) por lo general los segmentos deben tener un número mínimo de clientes que hagan rentables las inversiones en Marketing sobre dicho segmento.
- Accionables. Esto significa que dichos segmentos deben de tener capacidad de reacción ante las campañas de publicidad que se desarrollen.

Los individuos del mismo grupo deben de ser homogéneos entre sí, pero los distintos segmentos de mercado deben ser heterogéneos, deben tener entre sí diferencias significativas que generen reacciones diferentes ante diferencias en los productos y/o servicios. (Carrasco, 2020)

En este sentido, se podría decir que los clústeres formados con ambos modelos son cuantificables y ambos tienen un volumen mínimo de observaciones. De todas maneras, considero que tal vez los clústeres formados con k=4 son más accesibles. Al aumentar la cantidad de clústeres a 7 las campañas de marketing tal vez se vuelven un poco más ineficientes. Hay que pensar y planear 7 campañas diferentes en lugar de 4, lo cual consume tiempo y dinero. Si se segmentara a los clientes en 7 clústeres en lugar de 4, tal vez habría que identificar a cuáles segmentos amerita aplicarle una campaña y cuáles tal vez es mejor no dedicarles esfuerzo. Aquí surge entonces la siguiente pregunta: ¿Son más accionables los segmentos formados con k=7?

En este caso la única razón por la cual elegiría segmentar a mis clientes en 7 y no en 4 clústeres es si efectivamente los segmentos generados cuando k=7 son sustancialmente más accionables que cuando k=4. En este caso, no considero que sea así. A pesar de que se forman algunos segmentos interesantes cuando segmento a mis clientes en 7, no considero que los nuevos segmentos generados tengan una capacidad de reacción superior.

Se podría concluir entonces, que la mejor forma de segmentar a mis clientes teniendo en cuenta las tres dimensiones del modelo RFM es con k=4, formando los segmentos Vip, Nuevos, Churn y Peores descritos anteriormente.

Se van a observar 4 gráficos de tartas, uno para cada uno de los clústeres formados en función del RFM score de los clientes pertenecientes a ese clúster. Sería intuitivo que los clientes con RFM score más altos se encuentren en el clúster de clientes VIP y los clientes con RFM score más bajos en el clúster de los peores clientes. Asimismo, se realizó una tabla en donde se obtuvo el RFM score, mínimo, máximo y medio para cada clúster y se calculó el valor total que aporta cada clúster a la empresa:

Cluster RFM	Cantidad observaciones	RFM Score			
		Min	Max	Media	Total
1 (Vip)	464	3,33	5	4,29	1992
2 (Peores)	614	1	3	1,85	1136
3 (Churn)	518	2,66	4,33	3,47	1801
4 (Nuevos)	644	1,66	4	2,64	1698

Tabla 18 - RFM score para cada uno de los 4 segmentos encontrados

De la tabla, se puede observar que efectivamente el clúster que aporta más valor a la empresa es el de los clientes VIP. Incluso siendo el clúster con menos observaciones, es el clúster que, al sumar todos los scores de todos los clientes, alcanza un valor más

alto (1992). Por el contrario, los clientes pertenecientes al clúster 2, categorizados como los “peores clientes” son los que, sin duda, aportan menos valor a la empresa.

Si se observan los gráficos de tartas:

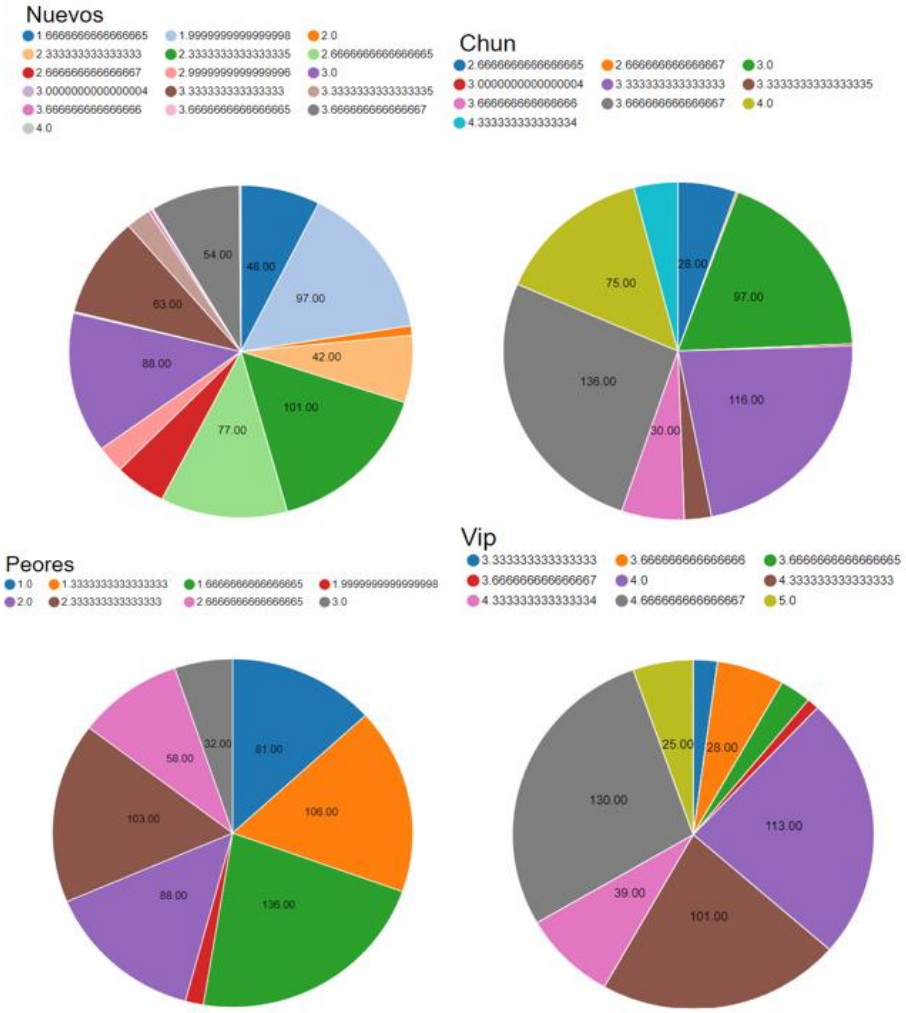


Ilustración 24 - Gráficos de tartas de los clústeres formados y el RFM score

Se puede observar que no hay ningún cliente Vip con un RFM score por debajo de 3. La mayoría de los clientes pertenecientes a este clúster tienen un puntaje superior a 4. Se puede observar también que todos los clientes con un RFM score de 5 pertenecen al clúster de los mejores clientes de la empresa.

Por el contrario, no hay ningún cliente considerado como de los peores clientes que tengan un puntaje superior a 3. Más de la mitad de los clientes no alcanzan siquiera a tener un puntaje 2. Asimismo, todos los clientes con un RFM score de 1 pertenecen a este clúster.

El clúster de los clientes considerados “nuevos”, es algo más variado, pero de todas maneras más de la mitad de los clientes no alcanzan a tener un RFM score de 3.

Por último, en el clúster de los clientes que están a punto de abandonar la empresa, considerados “Churn”, el RFM score tiende a ser más bien alto. La mayoría de los clientes alcanzan un RFM score de entre 3 y 4. Esto se debe a que son clientes con alta frecuencia y valor monetario pero que hace mucho no consumen en la empresa.

4.4.3. Segmentación de clientes en base a su ingreso, antigüedad y cantidad consumida en R.

En esta ocasión lo que se va a hacer es segmentar a los clientes en base a su ingreso, antigüedad y cantidad consumida. A diferencia de las dimensiones del modelo RFM, este tipo de segmentación no siempre se puede realizar, ya que no es tan fácil conseguir información acerca del ingreso de las personas. Es un dato que a mucha gente no le gusta brindar y por lo tanto no es tan fiable como son los datos que se pueden extraer con el consumo de los clientes. De todas maneras, en este caso tenemos la información, por lo que se va a utilizar para intentar sacar provecho de ella.

La segmentación de clientes en base a su ingreso, antigüedad y cantidad consumida se hará de la misma manera que se realizaron los clústeres con las 3 dimensiones del modelo RFM. Las variables ya normalizadas fueron llevadas a R.

En primer lugar, se va a observar la distribución de las variables:

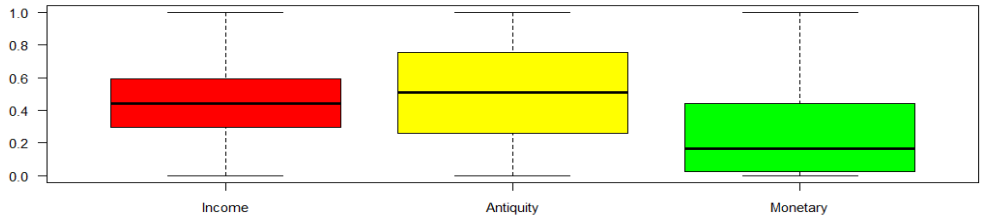


Ilustración 25 - Gráfico de cajas de las variables Ingreso, Antigüedad y Valor Monetario

Como era de esperarse, ya que fueron tratados en la etapa de depuración, ninguna de las tres variables parecería tener datos atípicos.

En segundo lugar, se puso a prueba el estadístico Hopkins. El mismo toma un valor de **0.798**. A pesar de que sea algo inferior al valor que toma el estadístico con las tres dimensiones del modelo RFM, sigue siendo bueno. Esto quiere decir que los datos no tienen una distribución aleatoria y podemos encontrar cúmulos en ellos.

Ahora voy a pasar a observar la distribución de los datos:

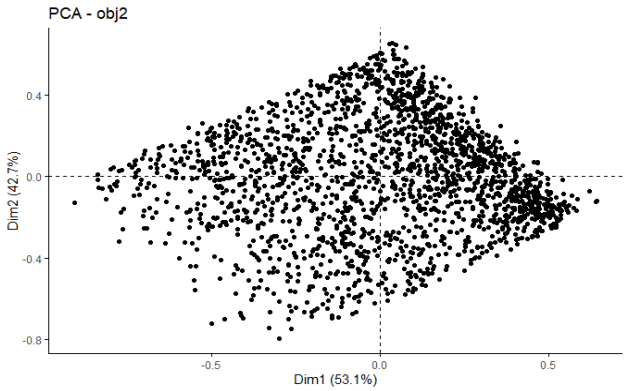


Ilustración 26 - Gráfico de dispersión Ingreso, Antigüedad y Valor Monetario

La forma es muy similar a lo que se observó en la etapa anterior. Se podrían dividir los datos en cuadrados.

Si se observa el mismo gráfico, pero ahora aplicando k-media para formar 4 clústeres distintos:

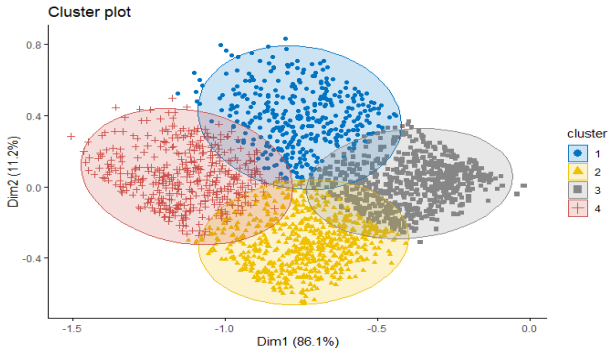


Ilustración 27 - Gráfico de dispersión de Ingreso, Antigüedad y Valor Monetario con visualización de 4 clústeres realizados con k-media

Los clústeres formados no son perfectamente circulares, pero también se podría decir que tienen una forma separable definida y por lo tanto que el algoritmo k-medias también funcionaría bien en estos datos. Hay algunos solapamientos en los clústeres y esos casos podrían ser objeto de estudio para entender mejor a qué grupo los deberíamos asignar.

Se va a intentar identificar la cantidad de clústeres óptima de la misma manera que en el caso anterior.

Gráfica Elbow:

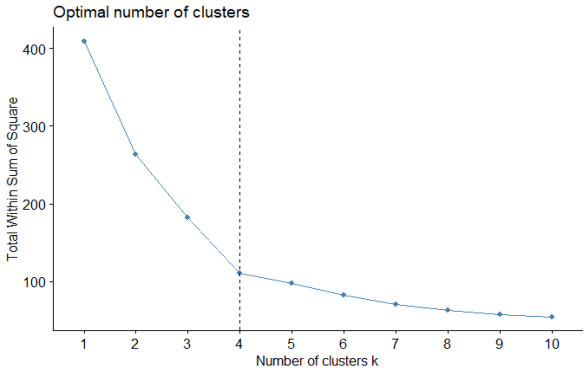


Ilustración 28 - Gráfica Elbow para clústeres con variables Ingreso, antigüedad y Valor Monetario

Gráfica Silhouette:

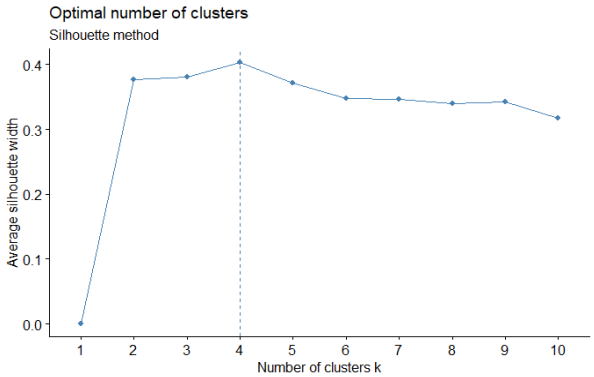


Ilustración 29 - Gráfica Silhouette para clústeres con variables Ingreso, antigüedad y Valor Monetario

Ambas encuentran el óptimo en k=4

Si se aplica la función NbClust de R y se corrobora la cantidad de clústeres óptima por 26 métodos distintos:

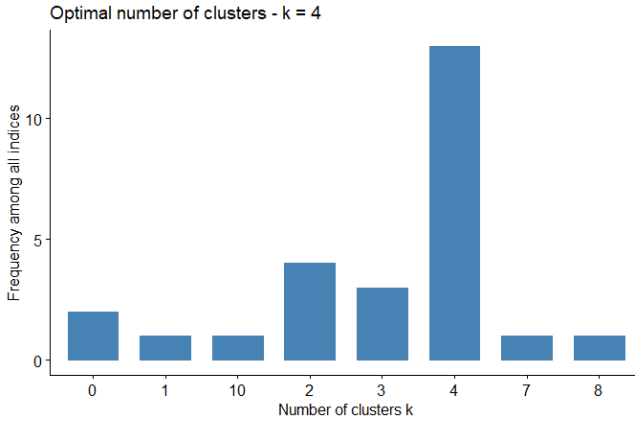


Ilustración 30 - Gráfica de barras número óptimo de clústeres

Se puede saber que de los 26 métodos utilizados para encontrar la cantidad de clústeres óptima:

- 13 propusieron 4 como el mejor número de clústeres.
- 4 propusieron 2 como el mejor número de clústeres.
- 3 propusieron 3 como el mejor número de clústeres.
- 2 propusieron 0 como el mejor número de clústeres.
- 1 propuso 1, 7, 8 y el 10 como el mejor número de clústeres.

Se podría concluir entonces que el número óptimo de clústeres es 4.

Si se analizan los clústeres formados cuando K=4:

	Income	Antiquity	Monetary	Cantidad Obs.	Perfil	Silueta
1	0.3403674	0.2443840	0.05992519	689	Atención	0,45
2	0.3181519	0.7349356	0.09450488	675	Perdidos	0,4
3	0.6418647	0.2605834	0.49635002	406	Potenciales	0,37
4	0.6124059	0.7719630	0.56126608	470	Estrella	0,37

Tabla 19 - Resumen clústeres en base al ingreso, antigüedad y valor monetario para k=4

Los 4 clústeres están relativamente bien formados. No hay ningún clúster con demasiadas observaciones, ni ninguno que tenga muy pocas. La silueta media en este caso es de 0.4. Asimismo, a pesar de que haya clústeres con una mejor silueta que otros, no hay ningún clúster que tenga una silueta por debajo de la media. Todos los clústeres formados parecerían ser medianamente buenos.

Clúster 1: Los clientes pertenecientes a este clúster fueron denominados clientes que “necesitan atención”. Son clientes relativamente nuevos, con ingreso por debajo de la media y gasto bajo. La razón por la cual necesitan atención es porque al ser nuevos en la empresa, se debe intentar que consuman hasta el máximo de su capacidad. Para que ello pase, se les debe ofrecer productos que sepamos que necesitan, ya que, al no tener ingresos altos, su gasto probablemente es destinado a productos de primera necesidad. Su capacidad de gasto es reducida, y, por lo tanto, el valor monetario difícilmente pueda llegar a ser similar al de los clientes pertenecientes a los clústeres 3 y 4, pero si al menos, logramos incrementar la frecuencia con la que consumen en la empresa, y los mantenemos activos, ya estamos ganando.

Clúster 2: En este clúster se encuentran los clientes “Perdidos”. Son clientes antiguos, con ingreso por debajo de la media y gasto bajo. Es probable que estos clientes hayan consumido muy pocas veces en la empresa y hace mucho tiempo. Difícilmente logremos realizar una campaña que llame su atención si todavía no hemos logrado que consuman. Son clientes con ingresos por debajo de la media, y, por ende, su capacidad de gasto es reducida. Se les llama “perdidos” no porque se sepa que ya no son clientes de la empresa, ya que esto no lo podemos saber hasta no observar el Recency, sino porque difícilmente logremos que incrementen su consumo, si todavía no lo hicieron después de tanto tiempo de estar vinculados con la empresa.

Clúster 3: Aquí se encuentran los clientes potenciales. Son clientes relativamente nuevos, con ingreso y gasto alto. No se les llama potenciales como sinónimo de *leads,* sino por el gran potencial de desarrollo que tienen como clientes, al tener ingreso y, por lo tanto, capacidad de gasto alto. Son clientes relativamente nuevos en comparación con el resto, pero en el tiempo que llevan vinculados con la empresa han logrado un gasto superior a la media. De todas maneras, aun no alcanzan a conseguir un valor monetario tan alto como los clientes estrella. Si logramos que sigan consumiendo, estos clientes se podrían convertir en los clientes estrella en poco tiempo. Hay que cuidarlos y fidelizarlos, intentando que perduren como clientes en el tiempo y que todos ellos formen parte de los clientes ‘VIP’ de la segmentación anterior.

Clúster 4: Clientes estrella. Son clientes antiguos con ingreso y gasto alto. Son seguramente los mejores clientes de la empresa, ya que han estado con nosotros hace mucho tiempo y queremos que sigan estando. Al tener ingresos altos, es probable que los clientes pertenecientes a este clúster respondan a ofertas de productos tanto de primera necesidad como productos que los consumen por placer. De todas maneras, en este clúster se podrían encontrar los clientes que fueron denominados Churn en la segmentación anterior. Al igual que los clientes en el clúster 3, hay que intentar que formen parte de los clientes ‘VIP’ de la empresa.

En este caso, no considero que se pueda llegar a formar otro segmento que pueda resultar interesante desde el punto de vista de negocio. Los cuatro clústeres formados brindan información suficiente. Asimismo, siendo que 4 es el número de clústeres que maximiza la silueta y añadir un clúster adicional apenas consigue mejorar la varianza interna, es probable que elija quedarme con esta segmentación.

Por mera curiosidad, voy a probar si al aumentar el número de clústeres, se forman segmentos que podrían ser relevantes desde el punto de vista de negocio.

K=5

	Income	Antiquity	Monetary	Cantidad Obs.	Perfil	Silueta
1	0.2979408	0.8257381	0.09426541	440	Perdidos	0,4
2	0.6134261	0.7728208	0.56292939	466	Estrella	0,37
3	0.3355934	0.1486670	0.04946263	428	Atención	0,45
4	0.6470138	0.2530828	0.50523613	389	Potenciales	0,37
5	0.3586474	0.4788078	0.09455258	517	Perdidos	0,28

Tabla 20 - Resumen clústeres en base al ingreso, antigüedad y valor monetario para k=5

En este caso, al observar los centroides de los clústeres formados, se puede ver que no se genera ningún clúster que sea relevante desde el punto de vista del negocio. A los 4 clústeres generados anteriormente, se adiciona un nuevo clúster de clientes con ingreso

por debajo del promedio, antigüedad media y valor monetario bajo. Estos clientes también fueron etiquetados como clientes perdidos, ya que es probable que hayan consumido en la empresa muy pocas veces y tienen ingreso y capacidad de gasto bajo. Asimismo, la silueta del nuevo clúster generado es muy baja, lo que significa que los clientes que se encuentran en ese clúster no son tan homogéneos entre sí.

4.4.4. Evaluación segmentación según ingreso, antigüedad y cantidad consumida

Del apartado anterior se pudo concluir que el número de clústeres óptimo para realizar esta segmentación es 4. De todas maneras, se va a estudiar en mayor profundidad los clústeres formados, para determinar si efectivamente la segmentación podría resultar de utilidad desde el punto de vista del negocio.

La forma de los clústeres para k=4 es la siguiente:

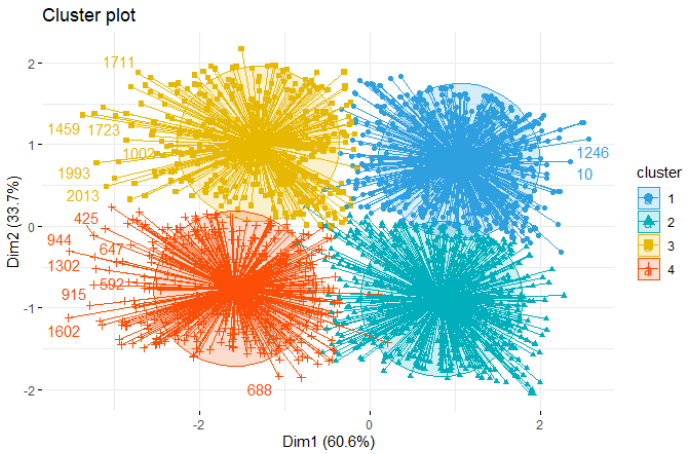


Ilustración 31 - observación clústeres en base al ingreso, antigüedad y valor monetario para k=4

A simple vista, parecerían estar bien formados.

Si se observa gráficamente la silueta de cada uno de los clústeres.

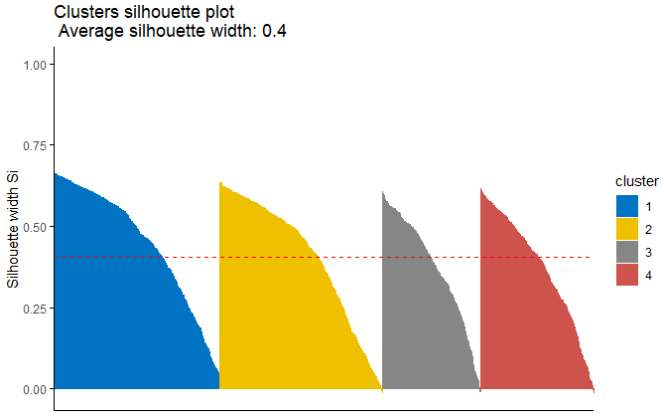


Ilustración 32 - Gráfica silueta media de clústeres en base al ingreso, antigüedad y valor monetario para k=4

Cuando k=4, el algoritmo k-media consigue formar clústeres con una silueta promedio de 0.4, lo cual es bueno. Asimismo, a pesar de que haya clústeres con una mejor silueta

que otros, no hay ningún clúster que tenga una silueta por debajo de la media. Todos los clústeres formados parecerían ser relativamente buenos.

Ahora voy a pasar a observar qué tan amplio es el rango de valores que toma cada variable en cada uno de los clústeres. Además, se agregó un resumen del RFM score que tiene cada uno de los clústeres:

Segunda Segmentación	Cantidad observaciones	Income			Antiquity			Monetary			RFM Score			
		Min	Max	Media	Min	Max	Media	Min	Max	Media	Min	Max	Media	Total
1 (Atención)	689	1730	87679	39858	185	526	356	6	985	146	1	4,33	2,24	1545
2 (Perdidos)	675	1730	92859	36443	524	884	698	5	839	225	1	4,66	2,45	1657
3 (Potenciales)	406	52845	113734	73450	185	555	367	277	2349	1156	2,33	5	3,89	1580
4 (Estrella)	470	2447	105471	70053	534	883	723	523	2352	1319	2,33	5	3,92	1844

Tabla 21 - Tabla con el mínimo, máximo y media de cada variable y del RFM score en cada clúster

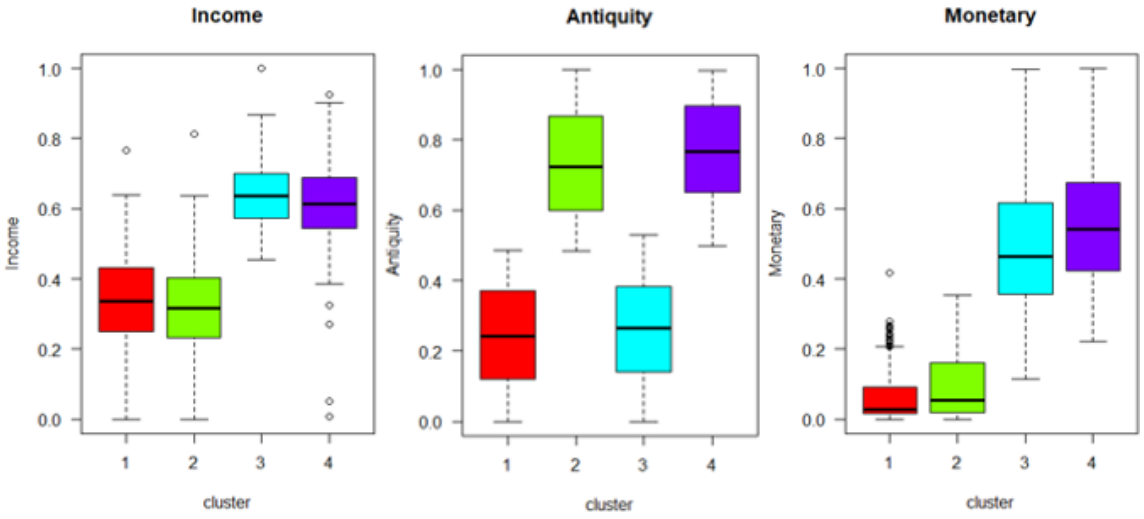


Ilustración 33 - Gráfico de cajas de las variables Income, Antiquity y Monetary para cada clúster

Se puede observar lo siguiente:

Tanto los clientes pertenecientes al clúster 1 (Atención) como al clúster 2 (Perdidos), tienen un ingreso medio / bajo y un gasto bajo. La diferencia entre ambos clústeres es que mientras los clientes del clúster 1 tuvieron un primer contacto con la empresa recientemente, los clientes del clúster 2 son clientes de la empresa hace más tiempo. Los clientes “Perdidos” fueron denominados así, ya que, a pesar de que hace mucho tiempo conocen la empresa, no han logrado incrementar su consumo. Ambos clústeres tienen clientes con RFM score de 1. Esto quiere decir que en ambos clústeres hay clientes con puntaje 1 en las 3 dimensiones del modelo RFM. Incluso el clúster “atención” presenta clientes con Recency score 1.

Por el contrario, tenemos los clientes pertenecientes al clúster 3 (Potenciales) y al clúster 4 (Estrella). Ambos tienen ingreso alto y gasto medio / alto. La diferencia entre ellos radica en la antigüedad. Mientras los clientes estrella son clientes de la empresa hace más tiempo, los clientes Potenciales consumen en la empresa hace menos tiempo. Los clientes potenciales han logrado tener un consumo muy alto en poco tiempo. Cabe destacar también, que en el clúster 4 hay algunos clientes con bajos ingresos, pero con alto consumo. Esto quiere decir que por más que no sea un patrón, una norma, podría pasar que clientes con bajos ingresos logren consumir más que la media.

En este sentido, se podría decir que hay dos grandes grupos de clientes, aquellos con ingreso alto y por ende gasto alto, y los clientes con ingreso y gasto más bien bajo. Dentro de los dos grupos, se puede diferenciar entre clientes nuevos, y clientes antiguos. Por otra parte, si se calcula el valor del cliente en base al RFM score, se puede ver que los clientes “Estrella” son los que aportan más valor a la empresa, incluso siendo que solo 470 clientes pertenecen a ese clúster. Además, a pesar de que el valor total que aportan los clústeres 1, 2 y 3 parezcan ser similares, en los clústeres 1 y 2 hay muchos más clientes que en el clúster 3, por lo tanto, en relación, el clúster 3 aporta mayor valor a la empresa.

Se van a tomar al azar 10 clientes de cada uno de los clústeres para ver de forma más clara si son homogéneos entre sí, y si los clústeres formados son heterogéneos:

Income	Antiquity	Monetary	cluster_obj4	Income	Antiquity	Monetary	cluster_obj4
58,138	848	1,617	4	63,564	336	1,215	3
82,800	767	1,315	4	83,443	365	1,497	3
76,995	643	1,782	4	58,330	371	1,064	3
2,447	724	1,730	4	75,507	243	1,440	3
58,607	738	972	4	58,512	468	1,171	3
79,632	679	637	4	79,146	251	564	3
68,657	679	1,196	4	78,285	429	1,427	3
48,948	698	902	4	75,127	223	833	3
80,011	611	1,395	4	72,063	546	758	3
72,550	783	1,319	4	78,939	384	1,507	3

Income	Antiquity	Monetary	cluster_obj4	Income	Antiquity	Monetary	cluster_obj4
37,758	818	40	2	23,661	392	23	1
7,500	721	15	2	7,500	370	57	1
38,683	818	341	2	60,182	228	22	1
57,906	568	401	2	22,804	518	26	1
43,456	645	393	2	31,686	209	17	1
53,172	604	486	2	31,160	471	64	1
30,545	805	69	2	29,938	430	26	1
42,835	549	595	2	40,737	388	17	1
39,922	685	156	2	44,159	197	275	1
17,117	713	128	2	22,070	453	67	1

Tabla 22 - Observaciones pertenecientes a cada clúster con k=4

Se puede observar cómo efectivamente los clientes pertenecientes a los clústeres 4 y 3 tienen un ingreso y un gasto superior a los clientes pertenecientes a los clústeres 2 y 1. Asimismo, también se puede notar que hay mayor diferencia en el gasto que en el ingreso. Con un ingreso no tanto superior se consume bastante más. Por otra parte, por mera casualidad se puede observar uno de los clientes atípicos del clúster 4 que mencionaba anteriormente. A pesar de tener ingresos por debajo de la media, tiene un consumo muy por arriba de la media. Cabe destacar que a pesar de que los clientes pertenecientes al clúster 3 y 1 efectivamente tengan una antigüedad más baja que los pertenecientes al clúster 4 y 2, los mismos no pueden ser considerados realmente nuevos, ya que hay algunos que llevan consumiendo en la empresa hace más de un año.

De todas maneras, considero que los clústeres formados cumplen con todas las características mencionadas anteriormente. Todos los clústeres son cuantificables, accesibles, tienen un volumen mínimo y son accionables. Los clientes dentro de cada clúster son homogéneos entre sí y los distintos segmentos son heterogéneos.

4.4.5. Primeras Conclusiones de las Segmentaciones

En este apartado me gustaría poder sacar algunas conclusiones con respecto a las dos segmentaciones realizadas, pero si se visualizan conjuntamente.

Podría ser interesante observar cuáles de los clientes pertenecientes al clúster 1 (Estrella) y 2 (Potenciales) pertenecen al clúster VIP de la segmentación realizada con las tres dimensiones del modelo RFM y cuáles pertenecen al clúster Churn.

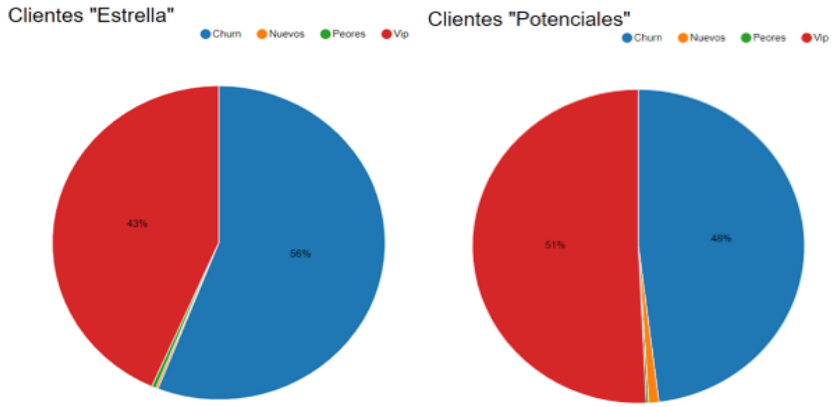


Ilustración 34 - Gráficos de tartas de los clientes "estrella" y "potenciales" en función de la segmentación RFM

De estos gráficos de tartas se pueden sacar algunas conclusiones. El 56% de los clientes “Estrella”, que son clientes antiguos con ingreso y gasto alto, son considerados “churn” en la segmentación realizada con las tres dimensiones del modelo RFM, mientras el 43% de los clientes son considerados VIP.

Por otra parte, a pesar de que los clientes “potenciales” hayan empezado a consumir en la empresa en promedio hace menos de un año, el 48% de ellos ya son considerados “churn” y el 51% ya son considerados “Vip”. Del total de los clientes “potenciales”, únicamente 4 son considerados “nuevos” en la segmentación anterior. De esta manera, es probable que los clientes considerados “potenciales” y “churn” hayan consumido mucho, en muy poco tiempo.

Son pocos los clientes que permanecen en el tiempo. Del total, únicamente 202 clientes son considerados al mismo tiempo “Estrella” y “Vip”. Por lo tanto, son sólo esos clientes los que realmente son fieles a la empresa. Es necesario realizar una campaña de fidelización para lograr una constancia de los clientes en el tiempo.

Si por otra parte se observan los clientes que necesitan atención o se consideran “Perdidos”:

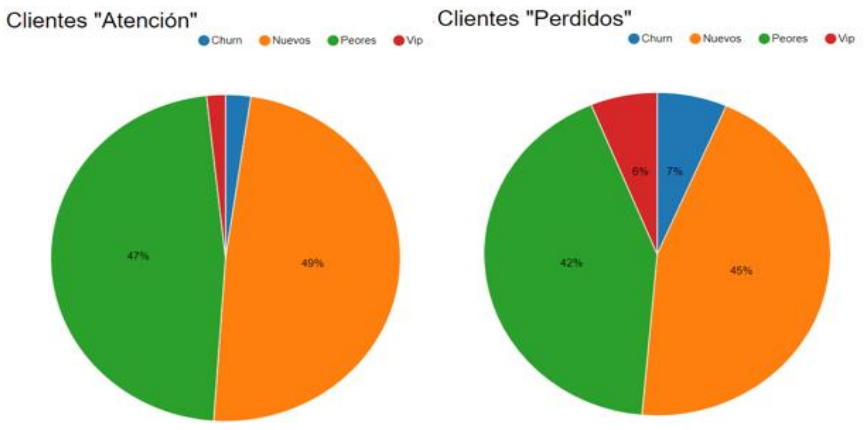


Ilustración 35 - Gráficos de tartas de los clientes "Atención" y "Perdidos" en función de la segmentación RFM

Del total de los clientes considerados como “necesitan atención” (principalmente porque se han incorporado recientemente a la empresa), el 47% de ellos ya se consideran parte

de “los peores” clientes según la segmentación RFM. Es probable que esos clientes hayan comprado una única vez en la empresa hace más de 1 mes.

Por otra parte, de los clientes “perdidos”, el 45% de ellos han consumido recientemente en la empresa, ya que también forman parte de los clientes “nuevos”. Estos clientes podrían ser clientes que en realidad no son nuevos, sino que se están reactivando. Consumieron hace mucho tiempo en la empresa, y lo volvieron a hacer recientemente. Los clientes considerados “perdidos” y “peores” al mismo tiempo, no sólo han consumido poco en la empresa, sino que son clientes que consumieron por última vez, hace mucho tiempo.

4.4.6. Algoritmo a priori

En este caso se va a aplicar el algoritmo a priori para determinar el perfil del consumidor de cada uno de los productos. En este sentido, se busca determinar los patrones de asociaciones más importantes para el conjunto de datos. La idea es intentar encontrar las características personales (edad, estado civil, educación, ingresos, etc.), de los mayores consumidores de vino, carne, pescado, dulces, frutas y productos de bazar, para poder aplicar las campañas de manera más eficiente. Una vez conocido el perfil del consumidor de vino, por ejemplo, las campañas vitivinícolas serán aplicadas principalmente a clientes con ese perfil.

Se van a intentar identificar por ejemplo las características (antecedentes) (X) -> del mayor consumidor de vino (consecuente) (Y).

Para poder aplicar el algoritmo de manera correcta, resulta interesante pasar las variables continuas a categóricas para que el algoritmo pueda buscar asociaciones más sencillas. El algoritmo lo que hace es buscar asociaciones entre ítems, por lo que al incluir en el algoritmo variables continuas con muchos posibles valores, su utilidad disminuye y la búsqueda de asociaciones es más difícil. Por esta razón, todas las variables continuas que se van a utilizar a la hora de aplicar el algoritmo se dividieron en intervalos. Para la preparación de los datos se utilizó la herramienta KNIME y para aplicar el algoritmo se utilizó la herramienta R.

Todas las variables relacionadas con el consumo del cliente fueron divididas en tres, con igual frecuencia (low, medium y high). Previo a la división en terciles, se excluyó a las personas que no habían consumido ese producto, creándoles la categoría “none”.

Las variables y sus categorías que van a ser consideradas para el algoritmo a priori serán las siguientes:

Cluster4: Es la segmentación realizada con las 3 dimensiones del modelo RFM con k=4. Las segmentaciones resultantes fueron, VIP, Churn, Nuevos y Peores.

Cluster_obj4: Es la segmentación realizada en base al ingreso, la antigüedad y valor monetario del cliente con k=4. Los posibles segmentos son, Estrella, Atención, Perdidos y Potenciales.

Educación: Grado de educación alcanzado por el cliente. Las categorías posibles son PhD, Master, Graduation y 2n Cycle.

Marital_Status: Esta variable nos dice si el cliente está soltero (Single) o en pareja (In Couple).

Edad: Edad del cliente. Se crearon los siguientes intervalos. Edades de entre 18 y 24 años, entre 25 y 34, entre 35 y 50, entre 50 y 65 y mayores de 65.

Teenhome: Esta variable toma la categoría “Teens” si hay adolescentes en el hogar y “No Teens” si no los hay.

Kidhome: Esta variable toma la categoría “Children” si hay niños en el hogar y “No Children” si no los hay.

Wine bin: Variable que indica el consumo de vino. Los intervalos son los siguientes:

- None wine: Clientes que nunca consumieron vino.
- Low wine: Clientes que gastaron entre 1 y 45 dólares en productos vitivinícolas.
- Médium wine: Clientes que gastaron entre 46 y 378 dólares en productos vitivinícolas.
- High wine: Clientes que gastaron entre 379 y 1493 dólares en productos vitivinícolas.

Fruits bin: Variable que indica el consumo de frutas. Los intervalos son los siguientes:

- None Fruits: Clientes que nunca consumieron frutas.
- Low Fruits: Clientes que gastaron entre 1 y 6 dólares en frutas.
- Médium Fruits: Clientes que gastaron entre 7 y 29 dólares en frutas.
- High Fruits: Clientes que gastaron entre 30 y 199 dólares en frutas.

Meet bin: Variable que indica el consumo de carne. Los intervalos son los siguientes:

- None Meet: Clientes que nunca consumieron carne.
- Low Meet: Clientes que consumieron entre 1 y 23 dólares en carne.
- Medium Meet: Clientes que consumieron entre 24 y 142 dólares en carne.
- High Meet: Clientes que consumieron entre 142 y 1725 dólares en carne.

Fish bin: Variable que indica el consumo de pescado. Los intervalos son los siguientes:

- None Fish: Clientes que nunca consumieron pescado.
- Low Fish: Clientes que consumieron entre 1 y 8 dólares en pescado.
- Medium Fish: Clientes que consumieron entre 10 y 40 dólares en pescado.
- High Fish: Clientes que consumieron entre 41 y 259 dólares en pescado.

Sweet bin: Variable que indica el consumo de dulces. Los intervalos son los siguientes:

- None Sweet: Clientes que nunca consumieron dulces.
- Low Sweet: Clientes que consumieron entre 1 y 6 dólares en dulces.
- Medium Sweet: Clientes que consumieron entre 7 y 29 dólares en dulces.
- High Sweet: Clientes que consumieron entre 30 y 263 dólares en dulces.

Gold bin: Variable que indica el consumo de productos de bazar. Los intervalos son los siguientes:

- None Gold: Clientes que nunca consumieron productos de bazar.
- Low Gold: Clientes que consumieron entre 1 y 13 dólares en bazar.
- Medium Gold: Clientes que consumieron entre 14 y 42 dólares en bazar.
- High Gold: Clientes que consumieron entre 43 y 362 dólares en bazar

Income bin: Variable que indica el ingreso del cliente. Los intervalos son los siguientes:

- Income low: Personas con ingresos entre 1,730 y 40,049.
- Income medium: Personas con ingresos entre 40,059 y 62,905.
- Income high: Personas con ingresos de entre 62,972 y 113,734.

Frequency bin: Variable que indica la frecuencia con la que el cliente realiza una compra. Los intervalos son los siguientes:

- Frequency Bad: Clientes que consumieron entre 0 y 6 veces en la empresa.
- Frequency Medium: Clientes que consumieron entre 7 y 16 veces en la empresa.
- Frequency Good: Clientes que consumieron entre 17 y 32 veces en la empresa.

Recency bin: Variable que indica el tiempo desde la última compra del cliente. Los intervalos son los siguientes:

- Recency Bad: Clientes que consumieron por última vez entre 66 y 99 días.
- Recency Medium: Clientes que consumieron por última vez entre 32 y 65 días.
- Recency Good: Clientes que consumieron por última vez entre 0 y 31 días.

Monetary bin: Variable que indica el gasto total en dólares del cliente. Los intervalos son los siguientes:

- Monetary Low: Clientes que gastaron en total entre 5 y 112 dólares.
- Monetary Medium: Clientes que gastaron en total entre 114 y 823 dólares.
- Monetary Good: Clientes que gastaron en total entre 825 y 2352 dólares.

Antiquity bin: Variable que indica el tiempo que hace que el cliente consume en la empresa. Los intervalos son los siguientes:

- Antiquity Low: Personas que son clientes de la empresa hace entre 185 y 428 días.
- Antiquity Medium: Personas que son clientes de la empresa hace entre 429 y 654 días.
- Antiquity high: Personas que son clientes de la empresa hace entre 655 y 884 días.

Si se observan gráficamente los 10 “ítems” con mayor frecuencia dentro de la base de datos (con ítems en este caso me refiero a todas las categorías de todas mis variables):

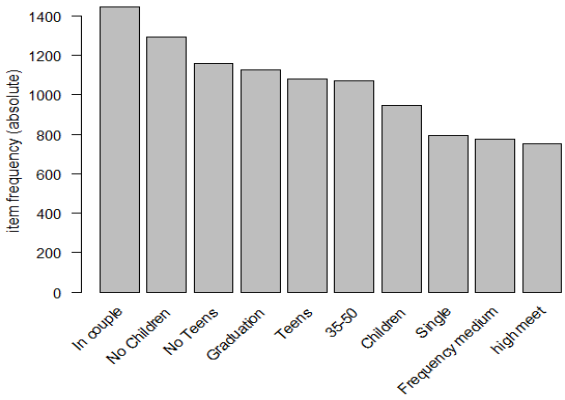


Ilustración 36 - Frecuencia de los 10 ítems más frecuentes

Dentro de la base de datos tenemos más clientes en pareja que solteros. El rango etario más frecuente son las personas entre 35 y 50 años, y el nivel educativo alcanzado que más se repite es “graduados”. Dado que esos son los *ítems* más frecuentes, es probable que haya más *itemsets* y luego reglas que estén formados con ellos. Asimismo, los 10 items graficados son los que presentan un soporte más alto, ya que, si recordamos, el soporte es la frecuencia relativa de los *ítems*.

Por otra parte, podemos saber que todas las categorías de las variables que eran de intervalo y fueron transformadas en categóricas, mediante la división de la variable en intervalos, van a tener aproximadamente el mismo soporte. Esto se debe a que a excepción de la variable Age, los intervalos fueron creados con igual frecuencia. En este sentido, el soporte de todas estas categorías debería ser de aproximadamente 0.33.

Se va a observar la frecuencia relativa de cada “*item*” para comprobar lo anterior:

```
> frecuencia_items %>% sort(decreasing = TRUE)
      In couple       No Children        No Teens       Graduation            Teens           35-50         Children          Single
      0.6446428571      0.5772321429      0.5169642857      0.5031250000      0.4830357143      0.4785714286      0.4227678571      0.3553571429
Frequency medium          high meet    Recency medium Antiquity  medium    Frequency good    Monetary  high       Income high        Income low
      0.3464285714      0.3361607143      0.3357142857      0.3343750000      0.3343750000      0.3343750000      0.3334821429      0.3334821429
  Antiquity  high     Income medium  Monetary  medium       Recency bad     Antiquity low       medium meet      Monetary low         high wine
      0.3330357143      0.3330357143      0.3330357143      0.3330357143      0.3325892857      0.3325892857      0.3325892857      0.3321428571
       medium wine      Recency good          low meet          low wine       medium gold         high gold          low gold    Frequency bad
      0.3316964286      0.3312500000      0.3308035714      0.3303571429      0.3294642857      0.3218750000      0.3214285714      0.3191964286
          Atencion          Perdidos             50-65             Nuevos          high fish         low fruits       medium fish     medium fruits
      0.3075892857      0.3013392857      0.2968750000      0.2875000000      0.2808035714      0.2808035714      0.2750000000      0.2745535714
             Peores         high sweet          low fish     medium sweet         low sweet        high fruits             Churn               PhD
      0.2741071429      0.2736607143      0.2727678571      0.2709821429      0.2683035714      0.2660714286      0.2312500000      0.2169642857
           Estrella               Vip         none sweet        Potenciales       none fruits         none fish            Master             25-34
      0.2098214286      0.2071428571      0.1870535714      0.1812500000      0.1785714286      0.1714285714      0.1651785714      0.1620535714
           2n Cycle               >65             18-24          none gold         none wine         none meet
      0.1147321429      0.0348214286      0.0276785714      0.0272321429      0.0058035714      0.0004464286
```

Ilustración 37 - Frecuencia relativa (soporte) de cada uno de los ítems pertenecientes a la base de datos

De esta forma podemos no sólo comprobar lo anterior, sino que también son muy pocos los *ítems* con frecuencia relativa menor a 1.

A diferencia de *“las canastas de compra”* para las cuales el algoritmo a priori se utiliza normalmente, en este caso, todas las transacciones de la base de datos a utilizar tienen la misma cantidad de *ítems.* Sabemos esto, ya que para poner a prueba el algoritmo, se va a utilizar la base de datos ya depurada, por lo que estamos seguros de que no tiene datos missing, lo que es lo mismo a decir que todas las transacciones tienen la misma cantidad de *ítems*.

Todas las transacciones tienen como mínimo y como máximo 18 *items*. En este sentido, es probable que se puedan formar varios *itemsets* con muchos *items*. Asimismo, hay 62 posibles *ítems,* pero cada transacción está compuesta por una categoría de cada variable, por lo tanto, una transacción no puede estar compuesta por cualquiera de los 62 *ítems* (una persona no puede estar soltero y en pareja a la vez).

Para aplicar el algoritmo se utilizará la función *apriori ()* de la librería “*arules*” de R. Esta función permite encontrar tanto *itemsets* frecuentes como reglas de asociación. Asimismo, permite determinar el nivel de soporte y confianza mínimo para la creación de itemsets y reglas, y es posible filtrar dependiendo del antecedente y el consecuente que yo quiero incluir. Es necesario crear un conjunto de datos transaccional para poder utilizar la función *apriori ().*

Para la creación de *itemsets* se exigirá un soporte mínimo de 0.1 y para la creación de reglas se exigirá un soporte mínimo de 0.1 y una confianza mínima de 0.6. Un soporte del 0.1 significa que cada *itemset* frecuente tiene que repetirse al menos 224 veces dentro de la base de datos (tenemos en total 2240 transacciones). Siendo que todas las

transacciones tienen 18 *ítems*, y las posibilidades de *ítems* dentro de las transacciones son acotadas, un soporte menor a 0.1 lo considero demasiado bajo. Asimismo, se pedirá que cada *itemset* formado al menos tenga dos *ítems*.

Hay casos que valores altos de confianza se deben a que el lado derecho de la regla tiene un soporte alto independiente del soporte del producto del lado izquierdo. Sabemos que el soporte de high wine, por ejemplo, es de 0.33, por lo tanto, la probabilidad de que un cliente pertenezca al intervalo de los mayores compradores de vino, dado determinado antecedente, debe ser obligatoriamente mayor a 0.33 para que dicha regla aporte información. Esto es también lo que da el Lift que debe ser obligatoriamente mayor a 1.

Si observamos un resumen de los *itemsets* frecuentes formados:

```
> summary(itemsets)
set of 3603 itemsets

most frequent items:
   No Children Monetary  high   Monetary low      high meet      In couple        (Other)
           785            683            651            636            598           9418

element (itemset/transaction) length distribution:sizes
   2    3    4    5    6    7
 694 1158 1019  570  151   11

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   3.000   3.000   3.545   4.000   7.000

summary of quality measures:
    support        transIdenticalToItemsets     count
 Min.   :0.1000   Min.   :0                Min.   :224.0
 1st Qu.:0.1080   1st Qu.:0                1st Qu.:242.0
 Median :0.1201   Median :0                Median :269.0
 Mean   :0.1317   Mean   :0                Mean   :295.1
 3rd Qu.:0.1437   3rd Qu.:0                3rd Qu.:322.0
 Max.   :0.3670   Max.   :0                Max.   :822.0

includes transaction ID lists: FALSE

mining info:
           data ntransactions support confidence
 caracteristicas          2240     0.1          1
```

Ilustración 38 - Resumen itemsets frecuentes

Podemos ver que se formaron 3603 *itemsets* frecuentes con un soporte mayor a 0.1. Los *ítems* más frecuentes dentro de estos *itemsets* formados son “No Children”, “Monetary high”, “Monetary low”, “High Meet” y “In Couple”. Asimismo, sabemos que el *itemset* con más *ítems* tiene 7 *items*. Por lo tanto, una regla a lo sumo va a estar compuesta por 6 antecedentes.

Si primero se observan los 20 *itemsets* más frecuentes dentro de la base de datos:

	items	support	transIdenticalToItemsets	coun
[1]	{In couple,No Children}	0.3669643	0	822
[2]	{In couple,No Teens}	0.3258929	0	730
[3]	{Graduation,In couple}	0.3209821	0	719
[4]	{In couple,Teens}	0.3187500	0	714
[5]	{35-50,In couple}	0.3160714	0	708
[6]	{Monetary high,No Children}	0.3133929	0	702
[7]	{Frequency bad,Monetary low}	0.3098214	0	694
[8]	{high meet,No Children}	0.3040179	0	681
[9]	{Income high,No Children}	0.3017857	0	676
[10]	{low wine,Monetary low}	0.3013393	0	675
[11]	{high wine,No Children}	0.3013393	0	675
[12]	{low meet,Monetary low}	0.3000000	0	672
[13]	{Frequency good,No Children}	0.3000000	0	672
[14]	{No Children,Teens}	0.2924107	0	655
[15]	{Graduation,No Children}	0.2901786	0	650
[16]	{high meet,Monetary high}	0.2892857	0	648
[17]	{Frequency bad,low meet}	0.2879464	0	645
[18]	{Frequency bad,low wine}	0.2857143	0	640
[19]	{No Children,No Teens}	0.2848214	0	638
[20]	{Frequency bad,low meet,Monetary low}	0.2830357	0	634

Tabla 23 - Itemsets frecuentes y su soporte

De los 20 itemsets formados con mayor soporte, tres de ellos podrían resultar interesantes desde el punto de vista del negocio. Se puede ver que se da de forma frecuente que los clientes con valor monetario alto, o con frecuencia alta, o que se encuentran en el intervalo de los clientes de mayor consumo de vino o carne, no tienen niños viviendo en el hogar.

Ahora voy a analizar las reglas encontradas con un soporte mínimo de 0.1 y una confianza mínima de 0.6, pero forzando que el consecuente sea high wine para intentar descubrir el perfil del cliente consumidor de vino.

4.4.6.1. Mayores consumidores de vino

Como ya fue advertido, a la hora de aplicar el algoritmo a priori, yo puedo obligar a que el algoritmo me dé únicamente aquellas reglas en donde el consecuente sea “high wine”. De esta forma puedo intentar encontrar las características que más se dan para que un cliente sea gran consumidor de vino. Si recordamos la ilustración 6, el vino es el producto más vendido de la empresa.

Si ordeno las reglas resultantes en base a la confianza, de forma descendiente:

```
> inspect(wines[1:30])
     lhs                                                                                rhs          support    confidence coverage   lift     count
[1]  {Monetary  high,Teens}                                                          => {high wine} 0.1107143 0.8920863  0.1241071 2.685851 248
[2]  {Monetary  high,No Children,Teens}                                              => {high wine} 0.1013393 0.8867188  0.1142857 2.669691 227
[3]  {Antiquity  high,Estrella,Monetary  high}                                       => {high wine} 0.1205357 0.8823529  0.1366071 2.656546 270
[4]  {Antiquity  high,Estrella,Monetary  high,No Children}                           => {high wine} 0.1098214 0.8817204  0.1245536 2.654642 246
[5]  {Antiquity  high,Monetary  high}                                                => {high wine} 0.1205357 0.8766234  0.1375000 2.639296 270
[6]  {Antiquity  high,Monetary  high,No Children}                                    => {high wine} 0.1098214 0.8754448  0.1254464 2.635748 246
[7]  {Frequency good,In couple,Income high,Monetary  high}                           => {high wine} 0.1160714 0.8724832  0.1330357 2.626831 260
[8]  {Frequency good,Income high,Monetary  high}                                     => {high wine} 0.1785714 0.8695652  0.2053571 2.618046 400
[9]  {Frequency good,In couple,Income high,Monetary  high,No Children}               => {high wine} 0.1080357 0.8673835  0.1245536 2.611477 242
[10] {Frequency good,high sweet,Income high,Monetary  high}                          => {high wine} 0.1223214 0.8670886  0.1410714 2.610589 274
[11] {Frequency good,high meet,In couple,Income high,Monetary  high}                 => {high wine} 0.1062500 0.8654545  0.1227679 2.605670 238
[12] {Frequency good,Income high,Monetary  high,No Children}                         => {high wine} 0.1687500 0.8649886  0.1950893 2.604267 378
[13] {Antiquity  high,Estrella,Frequency good}                                       => {high wine} 0.1022321 0.8641509  0.1183036 2.601745 229
[14] {Frequency good,high sweet,Income high,Monetary  high,No Children}              => {high wine} 0.1156250 0.8633333  0.1339286 2.599283 259
[15] {Estrella,Frequency good,Monetary  high}                                        => {high wine} 0.1294643 0.8605341  0.1504464 2.590855 290
[16] {Frequency good,high meet,high sweet,Income high,Monetary  high}                => {high wine} 0.1116071 0.8591065  0.1299107 2.586557 250
[17] {Frequency good,high fruits,Income high,Monetary  high}                         => {high wine} 0.1142857 0.8590604  0.1330357 2.586418 256
[18] {Estrella,Frequency good,Monetary  high,No Children}                            => {high wine} 0.1191964 0.8585209  0.1388393 2.584794 267
[19] {Frequency good,high meet,Income high,Monetary  high}                           => {high wine} 0.1593750 0.8581731  0.1857143 2.583747 357
[20] {Antiquity  high,Estrella,No Children}                                          => {high wine} 0.1160714 0.8580858  0.1352679 2.583484 260
```

Tabla 24 - Reglas de Asociación para mayores consumidores de vino

De la tabla anterior se puede decir, observando la regla 7, que con un 87% de seguridad, aquellas personas que tengan una frecuencia, un valor monetario y un ingreso alto, y, además, estén en pareja, van a pertenecer al intervalo de los mayores consumidores de vino. De todas formas, si a la regla 7 le quitamos el antecedente que además estén en pareja (regla 8), la confianza baja muy poco, pero el soporte sube. Y si a la regla 7 en lugar de que estar en pareja sea uno de los antecedentes, lo cambiamos por no tener niños en el hogar, la confianza también baja poco pero el soporte también sube (regla 12). El Lift es parecido en las 3 reglas. En estos casos, la seguridad con la que se da que, dado el antecedente, el cliente pertenezca al intervalo de los mayores consumidores de vino, es muy similar, pero la frecuencia con la que se dan las reglas 8 y 12 es superior a la frecuencia de la regla 7.

Para poder encontrar reglas que contengan la mayor cantidad de características posibles, se pueden observar las reglas maximales. Del marco teórico sabemos que para que un *itemset* se forme, todos los subconjuntos del *itemset* se tienen que dar. Se consideran reglas maximales cuando no hay otro *itemset* que sea *superset*. Es decir, es el *itemset* con más *items* posible.

Se van a observar las reglas maximales para ver si se obtienen reglas más interesantes desde el punto de vista del negocio:

```
> inspect(reglas_maximales[1:10])
     lhs                                                                                rhs          support    confidence coverage   lift     count
[1]  {Monetary  high,No Children,Teens}                                              => {high wine} 0.1013393 0.8867188  0.1142857 2.669691 227
[2]  {Antiquity  high,Estrella,Monetary  high,No Children}                           => {high wine} 0.1098214 0.8817204  0.1245536 2.654642 246
[3]  {Frequency good,In couple,Income high,Monetary  high,No Children}               => {high wine} 0.1080357 0.8673835  0.1245536 2.611477 242
[4]  {Frequency good,high meet,In couple,Income high,Monetary  high}                 => {high wine} 0.1062500 0.8654545  0.1227679 2.605670 238
[5]  {Antiquity  high,Estrella,Frequency good}                                       => {high wine} 0.1022321 0.8641509  0.1183036 2.601745 229
[6]  {Estrella,Frequency good,Monetary  high,No Children}                            => {high wine} 0.1191964 0.8585209  0.1388393 2.584794 267
[7]  {Frequency good,Monetary  high,Vip}                                             => {high wine} 0.1022321 0.8576779  0.1191964 2.582256 229
[8]  {Frequency good,high meet,high sweet,Income high,Monetary  high,No Children}    => {high wine} 0.1053571 0.8550725  0.1232143 2.574412 236
[9]  {Frequency good,high fruits,high meet,Income high,Monetary  high}               => {high wine} 0.1066964 0.8535714  0.1250000 2.569892 239
[10] {Income high,Monetary  high,No Children,Vip}                                    => {high wine} 0.1004464 0.8522727  0.1178571 2.565982 225
```

Tabla 25 - Reglas maximales para mayores consumidores de vino

En este caso, se puede decir con un 88% de seguridad que el perfil del cliente que más consume vino se encuentra en el intervalo de los clientes con valor monetario más alto, no tienen niños en el hogar, pero sí tienen adolescentes (regla 1). Asimismo, también se podría decir con un 86% de seguridad que, para ser parte del intervalo de los mayores consumidores de vino, el cliente debe tener una frecuencia y valor monetario alto, estar en pareja, tener ingresos entre 62,972 y 113,734 y no tener niños viviendo en el hogar (regla 3).

Por otra parte, existe lo que se llaman reglas redundantes. Se considera regla redundante si cubre la misma información, o información menos general, que la información que cubre otra regla de la misma utilidad y relevancia (misma o más confianza). En este sentido, es similar a “podar” la cantidad de reglas resultantes, dejando únicamente las más confiables sin perder información. Es probable que se muestren reglas que, con la menor cantidad de *ítems* posible, se maximice la confianza.

```
> inspect(wines_podado)
     lhs                                                          rhs         support   confidence coverage  lift     count
[1]  {Antiquity  high,Estrella,Monetary  high}                 => {high wine} 0.1205357 0.8823529  0.1366071 2.656546 270
[2]  {Antiquity  high,Monetary  high}                           => {high wine} 0.1205357 0.8766234  0.1375000 2.639296 270
[3]  {Frequency good,In couple,Income high,Monetary  high}     => {high wine} 0.1160714 0.8724832  0.1330357 2.626831 260
[4]  {Frequency good,Income high,Monetary  high}               => {high wine} 0.1785714 0.8695652  0.2053571 2.618046 400
[5]  {Antiquity  high,Estrella,Frequency good}                 => {high wine} 0.1022321 0.8641509  0.1183036 2.601745 229
[6]  {Estrella,Frequency good,Monetary  high}                  => {high wine} 0.1294643 0.8605341  0.1504464 2.590855 290
[7]  {Frequency good,Monetary  high,Vip}                       => {high wine} 0.1022321 0.8576779  0.1191964 2.582256 229
[8]  {Income high,Monetary  high,Vip}                          => {high wine} 0.1071429 0.8571429  0.1250000 2.580645 240
[9]  {Antiquity  high,Estrella}                                => {high wine} 0.1281250 0.8567164  0.1495536 2.579361 287
[10] {Estrella,Income high,Monetary  high}                     => {high wine} 0.1254464 0.8567073  0.1464286 2.579334 281
```

Tabla 26 - Reglas de Asociación para mayores consumidores de vino eliminando reglas redundantes

Eliminando las reglas redundantes, lo que se puede observar es que no hay ninguna regla que tenga como antecedente, pertenecer al intervalo de mayor consumidores de cualquier otro producto. Esto significa que los mayores consumidores de vino no consumen vino dado que consumen otro producto de la empresa. Primero consumen vino, y se verá luego si el consumo de vino es antecedente de algún otro producto.

Si ahora quisiéramos saber las características personales del consumidor de vino, forzando a que las posibilidades del “antecedente” sean únicamente las categorías "No Children", "Children", "No Teens", "Teens", "Income high", "Income medium", "Income low", "PhD", "Master", "Graduation", "2n Cycle", "Single", "In Couple", "18-24", "25-34, "35-50", "50-65", ">65", dejando de lado sus hábitos de compra, las reglas de asociación resultante son las siguientes:

```
> inspect(winesper)
    lhs                                          rhs         support   confidence coverage  lift     count
[1] {In couple,Income high,No Children}       => {high wine} 0.1459821 0.7622378  0.1915179 2.294909 327
[2] {Income high,No Children}                 => {high wine} 0.2290179 0.7588757  0.3017857 2.284787 513
[3] {Income high,No Children,No Teens}        => {high wine} 0.1482143 0.7494357  0.1977679 2.256365 332
[4] {In couple,Income high}                   => {high wine} 0.1593750 0.7437500  0.2142857 2.239247 357
[5] {Income high}                             => {high wine} 0.2468750 0.7402945  0.3334821 2.228844 553
[6] {Income high,No Teens}                    => {high wine} 0.1562500 0.7383966  0.2116071 2.223130 350
[7] {35-50,Income high}                       => {high wine} 0.1004464 0.7305195  0.1375000 2.199413 225
[8] {Graduation,Income high,No Children}      => {high wine} 0.1120536 0.7130682  0.1571429 2.146872 251
[9] {Graduation,Income high}                  => {high wine} 0.1223214 0.6936709  0.1763393 2.088471 274
```

Tabla 27 - Reglas de Asociación para mayores consumidores de vino, pero sin tener en cuenta su hábito de compra

En este caso, se forman únicamente 9 reglas que cumplen los requisitos de soporte mayor a 0.1 y confianza mayor a 0.6. Se podría decir que en el 76% de los casos en el que el cliente está en pareja, tiene un ingreso entre 62,972 y 113,734 y no tiene niños en el hogar, se encuentra en el intervalo de los mayores consumidores de vino. Esto

nos podría servir para identificar leads, ya que, al no depender de sus hábitos de compra para formar las reglas, es posible identificar perfiles externos a la empresa.

Si se repite lo mismo, pero para el resto de los productos de la empresa:

4.4.6.2. Mayores consumidores de carne

Si primero se observan las 20 reglas resultantes con mayor confianza cuando forzamos que el consecuente sea carne, con un soporte mínimo de 0.1 y una confianza mínima de 0.6:

```
> inspect(meet)
     lhs                                                                                                  rhs         support   confidence coverage  lift     count
[1]  {Graduation,high fruits,Monetary  high}                                                           => {high meet} 0.1138393 0.9550562  0.1191964 2.841070 255
[2]  {Graduation,high fish,high sweet,Monetary  high}                                                  => {high meet} 0.1008929 0.9535865  0.1058036 2.836698 226
[3]  {Graduation,high fruits,Monetary  high,No Children}                                               => {high meet} 0.1066964 0.9521912  0.1120536 2.832548 239
[4]  {Churn,high fruits,Monetary  high}                                                                => {high meet} 0.1053571 0.9516129  0.1107143 2.830827 236
[5]  {high fish,high fruits,Income high,Monetary  high}                                                => {high meet} 0.1294643 0.9508197  0.1361607 2.828468 290
[6]  {high fish,high fruits,high sweet,Income high,Monetary  high}                                     => {high meet} 0.1093750 0.9496124  0.1151786 2.824876 245
[7]  {high fruits,Income high,Monetary  high,No Children,No Teens}                                     => {high meet} 0.1174107 0.9494585  0.1236607 2.824418 263
[8]  {high fruits,Income high,Monetary  high,No Teens}                                                 => {high meet} 0.1227679 0.9482759  0.1294643 2.820900 275
[9]  {Graduation,high fish,Monetary  high}                                                             => {high meet} 0.1223214 0.9480969  0.1290179 2.820368 274
[10] {high fish,high fruits,Income high,Monetary  high,No Children}                                    => {high meet} 0.1223214 0.9480969  0.1290179 2.820368 274
[11] {high fish,high fruits,Monetary  high,No Teens}                                                   => {high meet} 0.1058036 0.9480000  0.1116071 2.820080 237
[12] {high fruits,high sweet,Monetary  high,No Teens}                                                  => {high meet} 0.1044643 0.9473684  0.1102679 2.818201 234
[13] {high fish,high fruits,high sweet,Income high,Monetary  high,No Children}                         => {high meet} 0.1035714 0.9469388  0.1093750 2.816923 232
[14] {high fish,high fruits,Monetary  high}                                                            => {high meet} 0.1500000 0.9464789  0.1584821 2.815555 336
[15] {high fish,high fruits,high gold,Monetary  high}                                                  => {high meet} 0.1022321 0.9462810  0.1080357 2.814966 229
[16] {high fruits,Monetary  high,No Children,No Teens}                                                 => {high meet} 0.1254464 0.9461279  0.1325893 2.814511 281
[17] {high fish,high fruits,high sweet,Monetary  high}                                                 => {high meet} 0.1245536 0.9457627  0.1316964 2.813424 279
[18] {high fruits,Monetary  high,No Teens}                                                             => {high meet} 0.1321429 0.9456869  0.1397321 2.813199 296
[19] {high fish,high fruits,Monetary  high,No Children,No Teens}                                       => {high meet} 0.1004464 0.9453782  0.1062500 2.812280 225
[20] {Graduation,high fish,Monetary  high,No Children}                                                 => {high meet} 0.1142857 0.9446494  0.1209821 2.810113 256
```

Tabla 28 - Reglas de Asociación para mayores consumidores de carne

Se puede observar que contrariamente a lo que sucedía con los mayores consumidores de vino, en este caso se da que, para ser parte del intervalo de los mayores consumidores de carne, primero deben ser parte de los mayores consumidores de algún otro producto vendido en la empresa. Por lo tanto, se podría decir que los grandes consumidores de carne no consumen únicamente carne. La mayoría de las reglas generadas con mayor confianza para los mayores consumidores de carne, tienen como antecedente que sean parte del intervalo de los mayores consumidores de fruta y pescado.

Si entonces ahora, al igual que con los mayores consumidores de vino, observamos las reglas maximales y las reglas “podadas”:

```
> inspect(reglas_maximalesmeet[1:10])
     lhs                                                                                                  rhs         support   confidence coverage  lift     count
[1]  {Graduation,high fish,high sweet,Monetary  high}                                                  => {high meet} 0.1008929 0.9535865  0.1058036 2.836698 226
[2]  {Graduation,high fruits,Monetary  high,No Children}                                               => {high meet} 0.1066964 0.9521912  0.1120536 2.832548 239
[3]  {Churn,high fruits,Monetary  high}                                                                => {high meet} 0.1053571 0.9516129  0.1107143 2.830827 236
[4]  {high fruits,Income high,Monetary  high,No Children,No Teens}                                     => {high meet} 0.1174107 0.9494585  0.1236607 2.824418 263
[5]  {high fruits,high sweet,Monetary  high,No Teens}                                                  => {high meet} 0.1044643 0.9473684  0.1102679 2.818201 234
[6]  {high fish,high fruits,high sweet,Income high,Monetary  high,No Children}                         => {high meet} 0.1035714 0.9469388  0.1093750 2.816923 232
[7]  {high fish,high fruits,high gold,Monetary  high}                                                  => {high meet} 0.1022321 0.9462810  0.1080357 2.814966 229
[8]  {high fish,high fruits,Monetary  high,No Children,No Teens}                                       => {high meet} 0.1004464 0.9453782  0.1062500 2.812280 225
[9]  {Graduation,high fish,Monetary  high,No Children}                                                 => {high meet} 0.1142857 0.9446494  0.1209821 2.810113 256
[10] {Graduation,high fish,Income high,Monetary  high}                                                 => {high meet} 0.1058036 0.9442231  0.1120536 2.808844 237
```

Tabla 29 - Reglas maximales para mayores consumidores de carne

```
> inspect(meet_podado)
     lhs                                                                                                  rhs         support   confidence coverage  lift     count
[1]  {Graduation,high fruits,Monetary  high}                                                           => {high meet} 0.1138393 0.9550562  0.1191964 2.841070 255
[2]  {Graduation,high fish,high sweet,Monetary  high}                                                  => {high meet} 0.1008929 0.9535865  0.1058036 2.836698 226
[3]  {Churn,high fruits,Monetary  high}                                                                => {high meet} 0.1053571 0.9516129  0.1107143 2.830827 236
[4]  {high fish,high fruits,Income high,Monetary  high}                                                => {high meet} 0.1294643 0.9508197  0.1361607 2.828468 290
[5]  {high fruits,Income high,Monetary  high,No Children,No Teens}                                     => {high meet} 0.1174107 0.9494585  0.1236607 2.824418 263
[6]  {high fruits,Income high,Monetary  high,No Teens}                                                 => {high meet} 0.1227679 0.9482759  0.1294643 2.820900 275
[7]  {Graduation,high fish,Monetary  high}                                                             => {high meet} 0.1223214 0.9480969  0.1290179 2.820368 274
[8]  {high fish,high fruits,Monetary  high,No Teens}                                                   => {high meet} 0.1058036 0.9480000  0.1116071 2.820080 237
[9]  {high fruits,high sweet,Monetary  high,No Teens}                                                  => {high meet} 0.1044643 0.9473684  0.1102679 2.818201 234
[10] {high fish,high fruits,Monetary  high}                                                            => {high meet} 0.1500000 0.9464789  0.1584821 2.815555 336
[11] {high fruits,Monetary  high,No Children,No Teens}                                                 => {high meet} 0.1254464 0.9461279  0.1325893 2.814511 281
[12] {high fruits,Monetary  high,No Teens}                                                             => {high meet} 0.1321429 0.9456869  0.1397321 2.813199 296
[13] {Graduation,high sweet,Monetary  high}                                                            => {high meet} 0.1205357 0.9440559  0.1276786 2.808347 270
[14] {high fruits,high sweet,Income high,Monetary  high}                                               => {high meet} 0.1325893 0.9428571  0.1406250 2.804781 297
[15] {high fruits,Income high,Monetary  high}                                                          => {high meet} 0.1669643 0.9420655  0.1772321 2.802426 374
```

Tabla 30 - Reglas de Asociación para mayores consumidores de carne, eliminando reglas redundantes

En este caso, al observar las reglas maximales y las reglas “podadas”, no se consigue nueva información. Lo único que resulta interesante observar, es que high wine, no es antecedente en ninguna regla cuando el consecuente es high Meet.

Si entonces se pasa, al igual que con los mayores consumidores de vino, a intentar identificar las características de los compradores de carne, independientemente del resto de los hábitos de compra del cliente:

```
> inspect(meetper)
     lhs                                                rhs          support    confidence coverage  lift     count
[1]  {Graduation,Income high,No Teens}               => {high meet} 0.1017857 0.9156627  0.1111607 2.723884 228
[2]  {In couple,Income high,No Children,No Teens}    => {high meet} 0.1093750 0.9074074  0.1205357 2.699326 245
[3]  {Income high,No Children,No Teens}              => {high meet} 0.1781250 0.9006772  0.1977679 2.679305 399
[4]  {In couple,Income high,No Teens}                => {high meet} 0.1165179 0.8877551  0.1312500 2.640865 261
[5]  {Income high,No Teens}                          => {high meet} 0.1875000 0.8860759  0.2116071 2.635870 420
[6]  {Graduation,Income high,No Children}            => {high meet} 0.1366071 0.8693182  0.1571429 2.586020 306
[7]  {Graduation,Income high}                        => {high meet} 0.1495536 0.8481013  0.1763393 2.522904 335
[8]  {In couple,Income high,No Children}             => {high meet} 0.1575893 0.8228438  0.1915179 2.447769 353
[9]  {Income high,No Children}                       => {high meet} 0.2482143 0.8224852  0.3017857 2.446702 556
[10] {Income high}                                   => {high meet} 0.2665179 0.7991968  0.3334821 2.377425 597
[11] {In couple,Income high}                         => {high meet} 0.1696429 0.7916667  0.2142857 2.355024 380
[12] {35-50,Income high}                             => {high meet} 0.1058036 0.7694805  0.1375000 2.289026 237
```

Tabla 31 - Reglas de Asociación para mayores consumidores de carne sin tener en cuenta los hábitos de compra

En este caso, los resultados son similares a los que se obtenían con los mayores consumidores de vino. Se logra una confianza más alta pero el soporte y el Lift son similares.

4.4.6.3. Mayores consumidores de fruta

Se observan las reglas creadas para los mayores consumidores de fruta:

```
> inspect(fruit)
     lhs                                                                    rhs           support    confidence coverage  lift     count
[1]  {high fish,high meet,high sweet,Income high}                        => {high fruits} 0.1156250 0.7848485  0.1473214 2.949766 259
[2]  {high fish,high meet,high sweet,Income high,No Children}             => {high fruits} 0.1089286 0.7820513  0.1392857 2.939253 244
[3]  {high fish,high meet,high sweet,Income high,Monetary  high}          => {high fruits} 0.1093750 0.7802548  0.1401786 2.932501 245
[4]  {high gold,high sweet,Monetary  high,No Children}                    => {high fruits} 0.1013393 0.7773973  0.1303571 2.921762 227
[5]  {high fish,high meet,high sweet,Income high,Monetary  high,No Children} => {high fruits} 0.1035714 0.7759197  0.1334821 2.916208 232
[6]  {high gold,high meet,high sweet,No Children}                         => {high fruits} 0.1004464 0.7758621  0.1294643 2.915992 225
[7]  {high gold,high meet,high sweet}                                     => {high fruits} 0.1084821 0.7738854  0.1401786 2.908562 243
[8]  {high gold,high sweet,Income high}                                   => {high fruits} 0.1013393 0.7721088  0.1312500 2.901886 227
[9]  {high gold,high sweet,Monetary  high}                                => {high fruits} 0.1066964 0.7709677  0.1383929 2.897597 239
[10] {high fish,high meet,high sweet,No Children}                         => {high fruits} 0.1276786 0.7708895  0.1656250 2.897303 286
[11] {high fish,high meet,high sweet,Monetary  high}                      => {high fruits} 0.1245536 0.7707182  0.1616071 2.896659 279
[12] {high fish,high meet,high sweet,Monetary  high,No Children}          => {high fruits} 0.1183036 0.7703488  0.1535714 2.895271 265
```

Tabla 32 - Reglas de Asociación para mayores consumidores de fruta

En este caso sucede algo muy similar a la carne. Para ser parte del intervalo de los mayores consumidores de fruta, también deben pertenecer al intervalo de los mayores consumidores de pescado, carne y dulces, o pertenecer al intervalo de los mayores consumidores de productos de bazar, carne y dulces además de no tener niños en el hogar, e ingresos altos.

Incluso si se observan las reglas obtenidas una vez quitadas las redundantes:

```
> inspect(fruit_podado)
     lhs                                                rhs           support    confidence coverage  lift     count
[1]  {high fish,high meet,high sweet,Income high}      => {high fruits} 0.1156250 0.7848485  0.1473214 2.949766 259
[2]  {high gold,high sweet,Monetary  high,No Children} => {high fruits} 0.1013393 0.7773973  0.1303571 2.921762 227
[3]  {high gold,high meet,high sweet,No Children}      => {high fruits} 0.1004464 0.7758621  0.1294643 2.915992 225
[4]  {high gold,high meet,high sweet}                  => {high fruits} 0.1084821 0.7738854  0.1401786 2.908562 243
[5]  {high gold,high sweet,Income high}                => {high fruits} 0.1013393 0.7721088  0.1312500 2.901886 227
[6]  {high gold,high sweet,Monetary  high}             => {high fruits} 0.1066964 0.7709677  0.1383929 2.897597 239
[7]  {high fish,high meet,high sweet,No Children}      => {high fruits} 0.1276786 0.7708895  0.1656250 2.897303 286
[8]  {high fish,high meet,high sweet,Monetary  high}   => {high fruits} 0.1245536 0.7707182  0.1616071 2.896659 279
[9]  {high fish,high meet,high sweet}                  => {high fruits} 0.1361607 0.7702020  0.1767857 2.894719 305
[10] {high fish,high sweet,Income high,Monetary  high} => {high fruits} 0.1151786 0.7655786  0.1504464 2.877343 258
```

Tabla 33 - Reglas de Asociación para mayores consumidores de fruta eliminando las reglas redundantes

Los antecedentes high Meet y high sweet aparecen en casi todas las reglas y en casi todas ellas también aparecen o high Gold o high fish. Se podría decir que, si se consume

carne, dulces y productos de bazar, o carne, dulces y pescado, también se va a consumir fruta. De nuevo, el único producto que no es antecedente de la fruta es el vino.

Si ahora se pasa a observar las características personales de los individuos que pertenecen al intervalo de los mayores consumidores de frutas, dejando de lado sus hábitos de compra:

```
> inspect(fruitper)
    lhs                                          rhs           support   confidence coverage  lift     count
[1] {Graduation,Income high}                  => {high fruits} 0.1156250 0.6556962  0.1763393 2.464362 259
[2] {Income high,No Teens}                    => {high fruits} 0.1383929 0.6540084  0.2116071 2.458018 310
[3] {Income high,No Children}                 => {high fruits} 0.1834821 0.6079882  0.3017857 2.285056 411
[4] {Graduation,Income high,No Children}      => {high fruits} 0.1066964 0.6789773  0.1571429 2.551861 239
[5] {Income high,No Children,No Teens}        => {high fruits} 0.1299107 0.6568849  0.1977679 2.468829 291
[6] {In couple,Income high,No Children}       => {high fruits} 0.1169643 0.6107226  0.1915179 2.295333 262
```

Tabla 34 - Reglas de Asociación para mayores consumidores de fruta sin tener en cuenta los hábitos de compra

A pesar de que la confianza es algo más baja que en el caso de la carne y el vino, las características personales de los individuos pertenecientes al intervalo de los mayores consumidores de frutas son similares a los mayores consumidores de vino y carne.

4.4.6.4. Mayores consumidores de pescado

Si se observan las reglas para los mayores consumidores de pescado:

```
> inspect(fish)
     lhs                                                                                rhs         support   confidence coverage  lift     count
[1]  {high gold,high meet,high sweet,Monetary  high,No Children}                     => {high fish} 0.1008929 0.8401487  0.1200893 2.991944 226
[2]  {high gold,high meet,high sweet,Monetary  high}                                  => {high fish} 0.1066964 0.8385965  0.1272321 2.986417 239
[3]  {Graduation,high meet,high sweet,Monetary  high}                                 => {high fish} 0.1008929 0.8370370  0.1205357 2.980863 226
[4]  {high gold,high sweet,Monetary  high}                                            => {high fish} 0.1151786 0.8322581  0.1383929 2.963844 258
[5]  {high gold,high sweet,Monetary  high,No Children}                                => {high fish} 0.1084821 0.8321918  0.1303571 2.963608 243
[6]  {high fruits,high meet,high sweet,Monetary  high,No Children}                    => {high fish} 0.1183036 0.8307210  0.1424107 2.958371 265
[7]  {high fruits,high meet,high sweet,Monetary  high}                                => {high fish} 0.1245536 0.8303571  0.1500000 2.957075 279
[8]  {Graduation,high sweet,Monetary  high}                                           => {high fish} 0.1058036 0.8286713  0.1276786 2.951071 237
[9]  {high fruits,high sweet,Monetary  high,No Children}                              => {high fish} 0.1254464 0.8264706  0.1517857 2.943234 281
[10] {high fruits,high sweet,Monetary  high}                                          => {high fish} 0.1316964 0.8263305  0.1593750 2.942735 295
```

Tabla 35 - Reglas de Asociación mayores consumidores de pescado

En este caso, sucede lo mismo que con la fruta.

```
> inspect(fish_podado)
     lhs                                                                                rhs         support   confidence coverage  lift     count
[1]  {high gold,high meet,high sweet,Monetary  high,No Children}                     => {high fish} 0.1008929 0.8401487  0.1200893 2.991944 226
[2]  {high gold,high meet,high sweet,Monetary  high}                                  => {high fish} 0.1066964 0.8385965  0.1272321 2.986417 239
[3]  {Graduation,high meet,high sweet,Monetary  high}                                 => {high fish} 0.1008929 0.8370370  0.1205357 2.980863 226
[4]  {high gold,high sweet,Monetary  high}                                            => {high fish} 0.1151786 0.8322581  0.1383929 2.963844 258
[5]  {high fruits,high meet,high sweet,Monetary  high,No Children}                    => {high fish} 0.1183036 0.8307210  0.1424107 2.958371 265
[6]  {high fruits,high meet,high sweet,Monetary  high}                                => {high fish} 0.1245536 0.8303571  0.1500000 2.957075 279
[7]  {Graduation,high sweet,Monetary  high}                                           => {high fish} 0.1058036 0.8286713  0.1276786 2.951071 237
[8]  {high fruits,high sweet,Monetary  high,No Children}                              => {high fish} 0.1254464 0.8264706  0.1517857 2.943234 281
[9]  {high fruits,high sweet,Monetary  high}                                          => {high fish} 0.1316964 0.8263305  0.1593750 2.942735 295
[10] {high fruits,high gold,high meet,Monetary  high}                                 => {high fish} 0.1022321 0.8237410  0.1241071 2.933513 229
```

Tabla 36 - Reglas de Asociación mayores consumidores de pescado eliminando las reglas redundantes

Para ser parte de los mayores consumidores de pescado, los clientes también deben ser parte de los mayores consumidores de otros tres productos entre productos de bazar, carne, fruta y dulces.

```
> inspect(fishper)
    lhs                                          rhs         support   confidence coverage  lift     count
[1] {Income high,No Children,No Teens}        => {high fish} 0.1459821 0.7381490  0.1977679 2.628702 327
[2] {Income high,No Teens}                    => {high fish} 0.1540179 0.7278481  0.2116071 2.592019 345
[3] {Graduation,Income high,No Children}      => {high fish} 0.1133929 0.7215909  0.1571429 2.569736 254
[4] {Graduation,Income high}                  => {high fish} 0.1227679 0.6962025  0.1763393 2.479322 275
[5] {Income high,No Children}                 => {high fish} 0.1986607 0.6582840  0.3017857 2.344287 445
[6] {In couple,Income high,No Children}       => {high fish} 0.1227679 0.6410256  0.1915179 2.282826 275
[7] {Income high}                             => {high fish} 0.2120536 0.6358768  0.3334821 2.264490 475
[8] {In couple,Income high}                   => {high fish} 0.1325893 0.6187500  0.2142857 2.203498 297
```

Tabla 37 - Reglas de Asociación mayores consumidores de pescado sin tener en cuenta los hábitos de compra

Las características personales de los consumidores de pescado son similares a las características encontradas en los otros productos.

4.4.6.5. Mayores consumidores de dulces

Si se observan las reglas para los mayores consumidores de dulces:

```
> inspect(sweet)
     lhs                                                                                  rhs          support   confidence coverage  lift     count
[1]  {high fish,high fruits,Income high,Monetary  high,No Children}                    => {high sweet} 0.1093750 0.8477509  0.1290179 3.097817 245
[2]  {high fish,high fruits,high meet,Income high,Monetary  high,No Children}          => {high sweet} 0.1035714 0.8467153  0.1223214 3.094033 232
[3]  {high fish,high fruits,Income high,Monetary  high}                                 => {high sweet} 0.1151786 0.8459016  0.1361607 3.091060 258
[4]  {high fish,high fruits,high meet,Income high,Monetary  high}                       => {high sweet} 0.1093750 0.8448276  0.1294643 3.087135 245
[5]  {high fish,high fruits,high meet,Income high,No Children}                          => {high sweet} 0.1089286 0.8442907  0.1290179 3.085173 244
[6]  {high fish,high fruits,high meet,Income high}                                      => {high sweet} 0.1156250 0.8436482  0.1370536 3.082825 259
[7]  {high fish,high fruits,Income high,No Children}                                    => {high sweet} 0.1169643 0.8424437  0.1388393 3.078424 262
[8]  {high fish,high fruits,Income high}                                                => {high sweet} 0.1241071 0.8373494  0.1482143 3.059809 278
[9]  {high fish,high fruits,Monetary  high,No Children}                                 => {high sweet} 0.1254464 0.8363095  0.1500000 3.056009 281
[10] {high fish,high fruits,high meet,Monetary  high,No Children}                       => {high sweet} 0.1183036 0.8359621  0.1415179 3.054739 265
[11] {high fish,high fruits,high wine,Monetary  high}                                   => {high sweet} 0.1022321 0.8357664  0.1223214 3.054024 229
[12] {high fish,high fruits,high wine}                                                  => {high sweet} 0.1040179 0.8321429  0.1250000 3.040783 233
```

Tabla 38 - Reglas de Asociación mayores consumidores de dulces

Al igual que sucedía con los últimos 3 productos analizados, para ser parte del intervalo de los mayores consumidores de dulces, también deben pertenecer al intervalo de los mayores consumidores de otros productos de la empresa. En este caso, la diferencia es que en las reglas 11 y 12, ser parte de los mayores consumidores de vino es antecedente de los mayores consumidores de dulces, y no hay ninguna regla en donde consumir productos de bazar, sea antecedente para consumir dulces.

Si se eliminan las reglas redundantes:

```
> inspect(sweet_podado)
     lhs                                                                                  rhs          support   confidence coverage  lift     count
[1]  {high fish,high fruits,Income high,Monetary  high,No Children}                    => {high sweet} 0.1093750 0.8477509  0.1290179 3.097817 245
[2]  {high fish,high fruits,Income high,Monetary  high}                                 => {high sweet} 0.1151786 0.8459016  0.1361607 3.091060 258
[3]  {high fish,high fruits,high meet,Income high,No Children}                          => {high sweet} 0.1089286 0.8442907  0.1290179 3.085173 244
[4]  {high fish,high fruits,high meet,Income high}                                      => {high sweet} 0.1156250 0.8436482  0.1370536 3.082825 259
[5]  {high fish,high fruits,Income high,No Children}                                    => {high sweet} 0.1169643 0.8424437  0.1388393 3.078424 262
[6]  {high fish,high fruits,Income high}                                                => {high sweet} 0.1241071 0.8373494  0.1482143 3.059809 278
[7]  {high fish,high fruits,Monetary  high,No Children}                                 => {high sweet} 0.1254464 0.8363095  0.1500000 3.056009 281
[8]  {high fish,high fruits,high wine,Monetary  high}                                   => {high sweet} 0.1022321 0.8357664  0.1223214 3.054024 229
[9]  {high fish,high fruits,high wine}                                                  => {high sweet} 0.1040179 0.8321429  0.1250000 3.040783 233
[10] {Frequency good,high fish,high fruits,No Children}                                 => {high sweet} 0.1035714 0.8315412  0.1245536 3.038585 232
[11] {high fish,high fruits,high meet,No Children}                                      => {high sweet} 0.1276786 0.8313953  0.1535714 3.038052 286
[12] {high fish,high fruits,Monetary  high}                                             => {high sweet} 0.1316964 0.8309859  0.1584821 3.036555 295
[13] {high fish,high fruits,high meet}                                                  => {high sweet} 0.1361607 0.8288043  0.1642857 3.028584 305
[14] {Churn,high fish}                                                                 => {high sweet} 0.1111607 0.8272425  0.1343750 3.022876 249
[15] {Frequency good,high fish,high fruits}                                            => {high sweet} 0.1111607 0.8272425  0.1343750 3.022876 249
```

Tabla 39 - Reglas de Asociación mayores consumidores de dulces eliminando las reglas redundantes

Al podar las reglas, high wine deja de ser antecedente de high sweet. En este sentido, ser consumidor de pescado y frutas es más importante para ser también consumidor de dulces que ser consumidor de vino.

Si se observan las reglas generadas dejando de lado los hábitos de compra:

```
> inspect(sweetper)
     lhs                                          rhs          support   confidence coverage  lift     count
[1] {Graduation,Income high,No Children}       => {high sweet} 0.1147321 0.7301136  0.1571429 2.667952 257
[2] {Graduation,Income high}                   => {high sweet} 0.1258929 0.7139241  0.1763393 2.608793 282
[3] {Income high,No Children,No Teens}         => {high sweet} 0.1401786 0.7088036  0.1977679 2.590082 314
[4] {Income high,No Teens}                     => {high sweet} 0.1495536 0.7067511  0.2116071 2.582581 335
[5] {Income high,No Children}                  => {high sweet} 0.1968750 0.6523669  0.3017857 2.383853 441
[6] {In couple,Income high,No Children}        => {high sweet} 0.1218750 0.6363636  0.1915179 2.325374 273
[7] {Income high}                              => {high sweet} 0.2120536 0.6358768  0.3334821 2.323596 475
[8] {In couple,Income high}                    => {high sweet} 0.1316964 0.6145833  0.2142857 2.245786 295
```

Tabla 40 - Reglas de Asociación mayores consumidores de dulces sin tener en cuenta los hábitos de compra

Las características personales dejando de lado los hábitos de compra, se repiten en este caso.

4.4.6.6. Mayores consumidores de productos de bazar

Si se observan las reglas para los mayores consumidores de productos de bazar:

```
> inspect(gold)
     lhs                                                    rhs         support   confidence coverage  lift     count
[1]  {Frequency good,Graduation,No Children}             => {high gold} 0.1080357 0.6974063  0.1549107 2.166699 242
[2]  {Graduation,Monetary  high,No Children}             => {high gold} 0.1133929 0.6902174  0.1642857 2.144365 254
[3]  {Graduation,Monetary  high}                         => {high gold} 0.1205357 0.6870229  0.1754464 2.134440 270
[4]  {Frequency good,Graduation}                         => {high gold} 0.1174107 0.6848958  0.1714286 2.127832 263
[5]  {high fish,high fruits,Monetary  high}              => {high gold} 0.1080357 0.6816901  0.1584821 2.117872 242
[6]  {high fish,high fruits,high meet,Monetary  high}    => {high gold} 0.1022321 0.6815476  0.1500000 2.117429 229
[7]  {Graduation,high meet,Monetary  high,No Children}   => {high gold} 0.1008929 0.6766467  0.1491071 2.102203 226
[8]  {Graduation,high meet,No Children}                  => {high gold} 0.1129464 0.6764706  0.1669643 2.101656 253
[9]  {high fish,high fruits,Monetary  high,No Children}  => {high gold} 0.1013393 0.6755952  0.1500000 2.098937 227
[10] {Graduation,high meet,Monetary  high}               => {high gold} 0.1080357 0.6740947  0.1602679 2.094275 242
[11] {Graduation,high fruits}                            => {high gold} 0.1026786 0.6705539  0.1531250 2.083274 230
[12] {high fruits,high sweet,Monetary  high}             => {high gold} 0.1066964 0.6694678  0.1593750 2.079900 239
```

Tabla 41 - Reglas de Asociación mayores consumidores de productos de bazar

Entre todos los productos, las reglas generadas con High Gold como consecuente son las que tienen sin duda la menor confianza, menor soporte y menor Lift. Asimismo, high wine no es antecedente en ninguna de las reglas generadas y high sweet únicamente en una. En este caso, ser consumidor de pescado, fruta o carne son antecedentes de ser consumidor de productos de bazar.

Si se eliminan las reglas redundantes:

```
> inspect(gold_podado)
     lhs                                                    rhs         support   confidence coverage  lift     count
[1]  {Frequency good,Graduation,No Children}             => {high gold} 0.1080357 0.6974063  0.1549107 2.166699 242
[2]  {Graduation,Monetary  high,No Children}             => {high gold} 0.1133929 0.6902174  0.1642857 2.144365 254
[3]  {Graduation,Monetary  high}                         => {high gold} 0.1205357 0.6870229  0.1754464 2.134440 270
[4]  {Frequency good,Graduation}                         => {high gold} 0.1174107 0.6848958  0.1714286 2.127832 263
[5]  {high fish,high fruits,Monetary  high}              => {high gold} 0.1080357 0.6816901  0.1584821 2.117872 242
[6]  {Graduation,high meet,No Children}                  => {high gold} 0.1129464 0.6764706  0.1669643 2.101656 253
[7]  {Graduation,high fruits}                            => {high gold} 0.1026786 0.6705539  0.1531250 2.083274 230
[8]  {high fruits,high sweet,Monetary  high}             => {high gold} 0.1066964 0.6694678  0.1593750 2.079900 239
[9]  {Graduation,high fish}                              => {high gold} 0.1084821 0.6694215  0.1620536 2.079756 243
[10] {Graduation,high meet}                              => {high gold} 0.1258929 0.6682464  0.1883929 2.076105 282
[11] {high fish,high sweet,Monetary  high}               => {high gold} 0.1151786 0.6632391  0.1736607 2.060549 258
[12] {Frequency good,high fish,Monetary  high}           => {high gold} 0.1116071 0.6631300  0.1683036 2.060210 250
```

Tabla 42 - Reglas de Asociación mayores consumidores de productos de bazar eliminando las reglas redundantes

Los resultados de las reglas una vez que se eliminan las reglas redundantes no cambian casi nada.

Si se observan las reglas generadas dejando de lado los hábitos de compra:

```
> inspect(golruper)
     lhs                                                rhs         support   confidence coverage  lift     count
[1] {Graduation,Income high,No Children} => {high gold} 0.1026786 0.6534091  0.1571429 2.030009 230
[2] {Graduation,Income high}             => {high gold} 0.1129464 0.6405063  0.1763393 1.989923 253
```

Tabla 43 - Reglas de Asociación mayores consumidores de productos de bazar sin tener en cuenta los hábitos de Compra

En este caso, se generan únicamente dos reglas con confianza mayor a 0.6 y soporte mayor a 0.1. En este sentido, se podría decir que el consumidor de productos de bazar no tiene un perfil tan marcado.

4.4.6.7. Clientes con mayor valor monetario

Para el caso de los clientes con mayor valor monetario, solo se van a mirar las características personales sin tener en cuenta los hábitos de compra. La razón de esto es que sería de esperarse que high wine y high Meet sean antecedentes de Monetary high, ya que son los dos productos que más se consumen en la empresa. Por lo tanto, tener en cuenta el consumo de cada producto no tiene tanto sentido.

La intención de este apartado pretende identificar las características personales de los mayores consumidores de la empresa independientemente de los productos que consuman.

```
> inspect(monetaryper)
     lhs                                                  rhs              support    confidence coverage  lift     count
[1]  {In couple,Income high,No Children,No Teens} => {Monetary  high} 0.1120536 0.9296296  0.1205357 2.780201 251
[2]  {Income high,No Children,No Teens}           => {Monetary  high} 0.1834821 0.9277652  0.1977679 2.774625 411
[3]  {Graduation,Income high,No Teens}            => {Monetary  high} 0.1013393 0.9116466  0.1111607 2.726420 227
[4]  {Income high,No Teens}                       => {Monetary  high} 0.1915179 0.9050633  0.2116071 2.706731 429
[5]  {In couple,Income high,No Teens}             => {Monetary  high} 0.1187500 0.9047619  0.1312500 2.705830 266
[6]  {Graduation,Income high,No Children}         => {Monetary  high} 0.1379464 0.8778409  0.1571429 2.625319 309
[7]  {Income high,No Children}                    => {Monetary  high} 0.2584821 0.8565089  0.3017857 2.561522 579
[8]  {Income high,Single}                         => {Monetary  high} 0.1008929 0.8464419  0.1191964 2.531415 226
[9]  {In couple,Income high,No Children}          => {Monetary  high} 0.1620536 0.8461538  0.1915179 2.530554 363
[10] {Graduation,Income high}                     => {Monetary  high} 0.1473214 0.8354430  0.1763393 2.498521 330
[11] {Income high}                                => {Monetary  high} 0.2741071 0.8219545  0.3334821 2.458182 614
[12] {In couple,Income high}                      => {Monetary  high} 0.1732143 0.8083333  0.2142857 2.417445 388
[13] {35-50,Income high}                          => {Monetary  high} 0.1084821 0.7889610  0.1375000 2.359510 243
[14] {Graduation,No Children,No Teens}            => {Monetary  high} 0.1031250 0.7196262  0.1433036 2.152153 231
[15] {In couple,No Children,No Teens}             => {Monetary  high} 0.1200893 0.7005208  0.1714286 2.095016 269
```

Tabla 44 - Reglas de Asociación clientes con mayor valor monetario sin tener en cuenta los hábitos de Compra

Las características personales de los mayores consumidores de la empresa en general son similares a las que se veían para cada producto en particular.

4.4.7. Evaluación y conclusiones Algoritmo A priori

Se puede decir entonces, que el consumo de vino es independiente del consumo de los otros productos, y que ser consumidor de vino no es antecedente del consumo de ningún otro producto de la empresa. En otras palabras, del análisis anterior, se pudo descubrir que, para ser consumidor de vino, no hay que ser consumidor de otros productos de la empresa, y ser consumidor de vino, no es antecedente de ser consumidor de ningún otro producto vendido en la empresa.

A los consumidores de vino, para que aumenten su consumo, se les podría aplicar campañas up selling, ofreciéndoles productos de mejor calidad, o cross selling, pero siempre dentro de los productos vitivinícolas. Si tenemos un cliente que consume vino tinto, tal vez se le puede ofrecer vino blanco.

La independencia de los productos vitivinícolas con el resto de los productos de la empresa llama un poco mi atención, dado que es, sin duda, el producto más consumido de la empresa. A priori, hubiera tendido a pensar que ser consumidor de vino iba a ser antecedente de ser consumidor de algún otro producto.

Por otra parte, sabemos que la empresa no se caracteriza por vender productos de bazar. No está muy definido el perfil de estos clientes, pero en el caso que sean grandes consumidores de productos de bazar, se les podría ofrecer otro tipo de productos. Por ejemplo, sabemos que muchas de las reglas generadas para los mayores consumidores de fruta y pescado, tienen como antecedente que pertenezcan al grupo de los mayores consumidores de productos de bazar. En este sentido, a los consumidores de productos de bazar, se les podría ofrecer fruta y pescado como campaña cross selling.

Dejando de lado los mayores consumidores de productos de bazar, las reglas en las que forzamos a que el consecuente sea high Fruits, son las que tienen una confianza más baja, y las reglas en las que forzamos que el consecuente sea high Meet son las que tienen una confianza más alta.

Siguiendo la regla 6 de la tabla 30, en un 94% de los casos, los clientes que son grandes consumidores de pescado, fruta y dulces, también van a ser grandes consumidores de

carne, siempre y cuando también sean parte de los clientes con mayores ingreso, mayor valor monetario y no tengan niños viviendo en el hogar. Por lo tanto, a estos clientes, también se les puede aplicar campañas cross selling.

Siguiendo la regla 1 de la tabla 34, con un 78% de seguridad, si el cliente pertenece al intervalo de los mayores consumidores de pescado, carne y dulces y, además, tiene ingresos altos, también va a ser parte de los mayores consumidores de frutas. Una vez más, en este caso se pueden aplicar campañas cross selling. Sí consumen pescado, carne y dulces, se les puede ofrecer frutas.

Lo mismo sucede con el pescado y con los dulces.

Siguiendo la regla 1 de la tabla 37, si el cliente es consumidor de productos de bazar, carne y dulces y, además, pertenece a los clientes con mayor valor monetario y no tiene niños en el hogar, con un 84% de seguridad, también va a ser consumidor de pescado. A los consumidores de productos de bazar, carne y dulces se les puede ofrecer también pescado.

Siguiendo la regla 4 de la tabla 40, si el cliente es consumidor de pescado, frutas y carne y además tiene ingresos altos, con un 84% de seguridad, también va a ser consumidor de dulces. A los consumidores de pescado, fruta y carne, se les puede ofrecer dulces.

Por último, me gustaría resaltar que las características personales de los consumidores de cada uno de los productos, dejando de lado los hábitos de consumo (tablas 27, 31, 34, 37, 40 y 43), son muy similares. Incluso el consumidor de vino presenta características similares al resto de los consumidores de la empresa. Siendo que el consumidor de vino es independiente del consumo del resto de los productos, podría pasar que las características personales fueran diferentes. En muchas reglas se repiten los ítems high income, No Children, No Teens, In couple, Graduation.

En este sentido, se podría decir, siguiendo la regla 1 de la tabla 45, que el perfil de los mayores consumidores de la empresa, independientemente de los productos que consuman es el siguiente:

- Están en pareja.
- Tienen ingresos entre 62,972 y 113,734.
- No tienen niños en el hogar
- No tienen adolescentes en el hogar

Las personas con las características mencionadas anteriormente, en un 92% de los casos, pertenecen al intervalo de mayor valor monetario.

5. Conclusiones

El desarrollo de este TFM me ha servido para profundizar más en los algoritmos de Machine Learning no supervisados. Gracias a ellos, se han identificado diferentes tipologías de clientes que van a ser la base de la toma de decisiones en el departamento de marketing.

Se ha podido cumplir con el principal objetivo del proyecto, alcanzando todos los objetivos secundarios. Luego de analizar, entender y preparar los datos, se han realizado dos segmentaciones distintas identificando, en cada una de ellas, 4 segmentos

diferentes. La primera de ellas basada en el modelo RFM nos ha dado cuatro tipos de clientes, los que hemos llamado Vip, peores, Churn y clientes nuevos, y en paralelo, la segunda segmentación para la que se han utilizado las variables asociadas con antigüedad como cliente, nivel de ingresos y nivel de gasto y con ellas se han definido otros cuatro segmentos que se han bautizado como Clientes Estrella, Perdidos, Potenciales y clientes que necesitan Atención.

Lo interesante de ambas segmentaciones ha sido además cruzarlas y observar cómo partiendo de pocas variables y gracias a algoritmos no supervisados, efectivamente hemos sido capaces de detectar información que era desconocida. Esto ha resultado extremadamente útil para el departamento de marketing, que va a poder definir sobre qué clientes merece la pena invertir, cómo fidelizar los que realmente son los mejores clientes, analizar las causas de su abandono en caso de que esto ocurra, etc.

Asimismo, con la utilización del algoritmo a priori, se han podido identificar las características de los mayores consumidores de la empresa, y se pudo detectar los productos que habitualmente se consumen de forma conjunta, y aquellos que se consumen solos. Lo interesante en este caso ha sido que el algoritmo apriori no se ha utilizado meramente para detectar asociaciones entre los productos consumidos por los clientes, sino que se ha aplicado para conseguir describir en más detalle características de los mejores consumidores de cada producto. De esta manera, se ha conseguido una herramienta muy interesante para aplicar en las campañas de marketing para captación de leads, y también para definir acciones de cross-selling y up-selling sobre la base actual de clientes de la compañía.

La forma en que el algoritmo a priori fue utilizado, podría resultar similar a aplicar un árbol de regresión de Machine Learning supervisado, realizando análisis predictivos. De esta manera, se puede ver que con la utilización de algoritmos no supervisados y un poco de imaginación, se puede realizar un proyecto muy enriquecedor y transformar los datos en información relevante para tomar mejores decisiones.

A pesar de suponer más esfuerzo, he desarrollado la preparación de los datos, así como los diferentes análisis, combinando tres herramientas: el software KNIME, que al ser un software open source y low code me ha resultado muy útil, y he visto que puede ser muy práctico para futuros proyectos; el software SAS Miner, ya que lo hemos ido utilizando durante el máster y partía de un conocimiento profundo previo sobre él, y, además, he enriquecido todo lo calculado aplicando código en R. La combinación de las tres herramientas me ha permitido afianzar conocimientos y comparar resultados. Me he dado cuenta de que las herramientas “low code” pueden facilitar mucho el trabajo en el momento de exploración y preparación de los datos, y de qué manera la adaptabilidad del código nos permite obtener resultados realmente exhaustivos, sobre todo en la parte de modelado y visualización.

De todas maneras, a pesar de que, a nivel personal y profesional, el TFM me fue de gran utilidad para afianzar mis conocimientos en las herramientas SAS Miner, KNIME y R, en un futuro me gustaría incursionar en la herramienta Python, que sé que es muy útil en todo lo que respecta a la minería de datos.

Asimismo, me sería interesante poder aplicar los conocimientos adquiridos en datos nuevos, de una empresa conocida, para poder poner en práctica los modelos

encontrados. En este sentido, me gustaría poder aplicar la fase 6 del modelo CRISP DM, implantando los modelos generados en una base de datos real para poder analizar realmente el impacto que tiene la minería de datos sobre las campañas de marketing. Demostrando, en base a resultados reales, cómo el estudio y análisis de los datos puede proporcionar un impulso significativo a la eficiencia de una campaña de marketing, aumentando las respuestas o reduciendo los gastos.

6. Bibliografía

Ansari, O. B. (16 de Junio de 2021). Geo-Marketing Segmentation with Deep Learning. *Businesses (MDPI)*, págs. 6-7.

Calviño Martinez, A. (2020). *Apuntes de clase asignatura SEMMA.* Facultad de Estudios Estadísticos: Universidad Complutense de Madrid.

Carrasco, R. (2020). *Apuntes de clase asignatura CRM.* Facultad de Estudios Estadísticos : Universidad Complutense de Madrid.

Chapman, P., Clinton, J., Kreber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (Agosto de 2000). *CRISP-DM 1.0 - Step-by-step data mining guide.* SPSS. Obtenido de https://www.the-modeling-agency.com/crisp-dm.pdf

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software*. Obtenido de https://www.rdocumentation.org/packages/NbClust/versions/3.0/topics/NbClust

Córdoba, G. (02 de Febrero de 2011). *Unica 360*. Obtenido de https://www.unica360.com/analisis-rfm-en-retail-empezando-a-segmentar-clientes-i

Corrales, J. A. (19 de Agosto de 2020). *Rock Content*. Obtenido de https://rockcontent.com/es/blog/segmentacion-de-clientes/

Das, S., & Cakmak, U. (2018). Hands-On Automated Machine Learning. En *Hands-On Automated Machine Learning* (págs. 35-36). Brimingham: Packt Publishing Ltd.

Esteban, P. G. (11 de 11 de 2020). *Blog Visionarios*. Obtenido de Blog Visionarios: https://blogvisionarios.com/e-learning/articulos-ia/que-es-el-clustering-segmenta-a-tus-clientes-con-machine-learning/

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. Massachusetts: ELSEVIER.

Kassambara, A. (2017). *Practical Guide To Cluster Analysis in R.* STHDA.

Kassambara, A. (2018). *Data Novia*. Obtenido de https://www.datanovia.com/en/lessons/assessing-clustering-tendency/

Kotler, P., & Armstrong, G. (2012). Marketing. En P. Kotler, & G. Armstrong, *Marketing* (pág. 11). Mexico: Pearson Educación.

Kumar, V., & Reinartz, W. (2018). *Customer Relationship Management.* Berlin: Springer-Verlag GmbH.

Martinez, R. G. (2021). *Apuntes de clase de la asignatura Modelos de Desición de Marketing.* Facultad de Estudios Estadísticos : Universidad Complutense de Madrid.

Martinez, R. G., Carrasco, R. A., Garcia-Madariaga, J., Porcel Gallego, C., & Herrera-Viedma, E. (2020). *A comparison between Fuzzy Linguistic RFM Model and traditional RFM Model applied to Campaing Management. Case study of retail business.* Elsevier B. V.

Martinez, R., Carrasco, R., Sanchez-Figueroa, C., & Gavilan, D. (2021 de Agosto de 2021). An RFM Model Customizable to Product Catalogues and Marketing Criteria Using Fuzzy Linguistic Models: Case Study of a Retail Business. *Mathematics*, págs. 2-3.

Moine, J. M., Haedo , A., & Gordillo, S. (24 de 06 de 2011). *Estudio comparativo de metodologías para minería de datos.* Buenos Aires. Obtenido de https://core.ac.uk/download/pdf/301040544.pdf

Naranjo Cuervo, R., & Sierra Martínez, L. (2009). Herramienta software para el análisis de canasta de mercado sin selección de candidatos . *INGENIERÍA E INVESTIGACIÓN VOL. 29*, 61-64.

Parvaneh, A., Abbasimehr, H., & Tarokh, M. (2012). Integrating AHP and Data Mining for Effective Retailer Segmentation Based on Retailer Lifetime Value. *Journal of Optimization in Industrial Engineering*, 25.

Pérez López, C. (2004). Técnicas de Análisis Multivariabte de Datos. Madrid: Pearson Educación S.A.

Peter, J., & Olson, J. (2006). Comportamiento del Consumidor y Estrategia de Marketing. México, D. F.: McGRAW-HILL/INTERAMERICANA EDITORES, S.A.

Pitol, F. (4 de Mayo de 2014). *Blog de Inteligencia Artificial en Español*. Obtenido de http://ferminpitol.blogspot.com/2014/05/reglas-de-asociacion-algoritmo-apriori.html

Portillo, R. (10 de 05 de 2021). *Marketing Insider Review*. Obtenido de https://www.marketinginsiderreview.com/importancia-analisis-datos-marketing/

Rodó, P. (6 de Julio de 2019). *Economipedia*. Obtenido de https://economipedia.com/definiciones/normalizacion-estadistica.html

Rodrigo, J. A. (Septiembre de 2017). *Ciencia de Datos*. Obtenido de https://www.cienciadedatos.net/documentos/37_clustering_y_heatmaps

Rodrigo, J. A. (Junio de 2018). *Ciencia de Datos*. Obtenido de https://www.cienciadedatos.net/documentos/43_reglas_de_asociacion

Rodrigo, J. A. (Octubre de 2020). *Ciencia de Datos* . Obtenido de https://www.cienciadedatos.net/documentos/py06_machine_learning_python_scikitlearn.html

Saldanha, R. (23 de 06 de 2021). *Kaggle*. Obtenido de https://www.kaggle.com/rodsaldanha/arketing-campaign

SAS. (23 de 06 de 2021). Obtenido de SAS: https://www.sas.com/es_es/insights/analytics/machine-learning.html

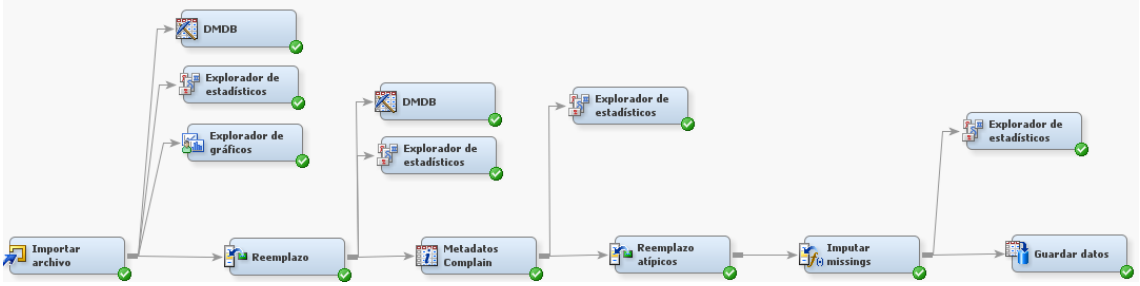
Weinstein, A. (2004). Handbook of Market Segmentation : Strategic Targeting for Business and Technology Firms. Taylor & Francis Group .

Wright, G. (10 de Agosto de 2021). *TechTarget*. Obtenido de https://searchdatamanagement.techtarget.com/definition/RFM-analysis

7. Anexo

7.1. Capturas configuración de los nodos de la depuración de datos SAS Miner

Primero que nada, voy a mostrar el flujo de la depuración:



La siguiente imagen, es la configuración del nodo explorador de estadísticos:

Entrenamiento	
Variables	
Datos	
Número de observaciones	100000
Validación	No
Prueba	No
Informes estándar	
Distribuciones del intervalo	Sí
Distribuciones de clase	Sí
Resumen de nivel	Sí
Utilizar variables de segmento	No
Tabulación cruzada	
Selección de variables	
Ocultar variables rechazadas	Sí
Número de variables seleccionadas	1000
Estadísticos chi-cuadrado	
Chi-cuadrado	Sí
Variables de intervalo	No
Número de clases	5
Estadísticos de correlación	
Correlaciones	Sí
Correlaciones Pearson	Sí
Correlaciones Spearman	No

La siguiente es la configuracion del nodo reemplazo utilizado para tratar los datos erroneos y luego los datos atípicos:

Entrenamiento	
Variables de intervalo	
Editor de reemplazo	
Método de límites predeterminado	Ninguno
Valores de corte	
Variables de clase	
Editor de reemplazo	
Niveles desconocidos	Ignorar
Puntuación	
Valores de sustitución	Calculado
Ocultar	No
Informe	
Informes de sustitución	Sí

Editor de reemplazo para las variables de intervalo (en esta ocasión se trataron los datos erróneos):

Nombre	Usar	Método de límite	Límite superior de reemplazo	Límite inferior de reemplazo	Método de sustitución
Age	Predetermir	Especificado por el usu	100	.	Ausente
Antiquity	Predetermir	Predeterminado	.	.	Predeterminado
Income	Predetermir	Especificado por el usu	666665	.	Ausente
MntFishProducts	Predetermir	Predeterminado	.	.	Predeterminado
MntFruits	Predetermir	Predeterminado	.	.	Predeterminado
MntGoldProds	Predetermir	Predeterminado	.	.	Predeterminado
MntMeatProducts	Predetermir	Predeterminado	.	.	Predeterminado
MntSweetProduc	Predetermir	Predeterminado	.	.	Predeterminado
MntWines	Predetermir	Predeterminado	.	.	Predeterminado
NumCatalogPurcl	Predetermir	Predeterminado	.	.	Predeterminado
NumDealsPurcha	Predetermir	Predeterminado	.	.	Predeterminado
NumStorePurcha	Predetermir	Predeterminado	.	.	Predeterminado
NumWebPurchas	Predetermir	Predeterminado	.	.	Predeterminado
NumWebVisitsMo	Predetermir	Predeterminado	.	.	Predeterminado
Recency	Predetermir	Predeterminado	.	.	Predeterminado

Editor de reemplazo para agrupar las variables de clase que no llegaban a tener un 5% de las observaciones:

Variable	Valor formateado	Valor de reemplazo	Número de ocurrencias	Variable	Valor formateado	Valor de reemplazo	Número de ocurrencias
Complain	1	**0**	21	Marital_Status	Divorced	**Single**	232
Complain	_UNKNOWN_	**_DEFAULT_**	.	Marital_Status	Widow	**Single**	77
Complain	0		2219	Marital_Status	Alone	**Single**	3
Education	Basic	**2n Cycle**	54	Marital_Status	Absurd	**Single**	2
Education	_UNKNOWN_	**_DEFAULT_**	.	Marital_Status	YOLO	**Single**	2
Education	Graduation		1127	Marital_Status	Married	**In couple**	864
Education	PhD		486	Marital_Status	Together	**In couple**	580
Education	Master		370	Marital_Status	_UNKNOWN_	**_DEFAULT_**	.
Education	2n Cycle		203	Marital_Status	Single		480
Kidhome	2	**1**	48	Teenhome	2	**1**	52
Kidhome	_UNKNOWN_	**_DEFAULT_**	.	Teenhome	_UNKNOWN_	**_DEFAULT_**	.
Kidhome	0		1293	Teenhome	0		1158
Kidhome	1		899	Teenhome	1		1030

Editor de reemplazo para las variables de intervalo, pero esta vez para tratar atípicos:

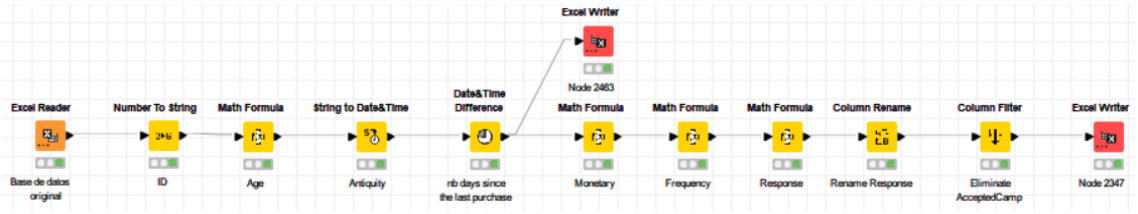
Nombre	Usar	Método de límite	Límite inferior de reemplazo	Límite superior de reemplazo	Método de sustitución
Antiquity	**Predetermir**	**Desviación estándar**	.	.	**Ausente**
Recency	**Predetermir**	**Desviación estándar**	.	.	**Ausente**
REP_Income	**Predetermir**	**Desviación estándar**	.	.	**Ausente**
Monetary	**Predetermir**	**Desviación estándar**	.	.	**Ausente**
Frequency	**Predetermir**	**Desviación estándar**	.	.	**Ausente**
Income	**Predetermir**	**Desviación estándar**	.	.	**Ausente**
MntFishProducts	**Predetermir**	**Predeterminado**	.	.	**Ausente**
MntGoldProds	**Predetermir**	**Predeterminado**	.	.	**Ausente**
Age	**Predetermir**	**Predeterminado**	.	.	**Ausente**
MntMeatProducts	**Predetermir**	**Predeterminado**	.	.	**Ausente**
MntSweetProduc	**Predetermir**	**Predeterminado**	.	.	**Ausente**
NumStorePurcha	**Predetermir**	**Predeterminado**	.	.	**Ausente**
MntWines	**Predetermir**	**Predeterminado**	.	.	**Ausente**
NumDealsPurcha	**Predetermir**	**Predeterminado**	.	.	**Ausente**
NumCatalogPurc	**Predetermir**	**Predeterminado**	.	.	**Ausente**
NumWebVisitsMo	**Predetermir**	**Predeterminado**	.	.	**Ausente**
NumWebPurchas	**Predetermir**	**Predeterminado**	.	.	**Ausente**
MntFruits	**Predetermir**	**Predeterminado**	.	.	**Ausente**
REP_Age	**Predetermir**	**Predeterminado**	.	.	**Ausente**
Response	**Predetermir**	**Predeterminado**	.	.	**Ausente**

Por último, una captura que muestra la forma que fueron imputados los datos missing. Se puede ver que se utilizó distribución como método de imputación.

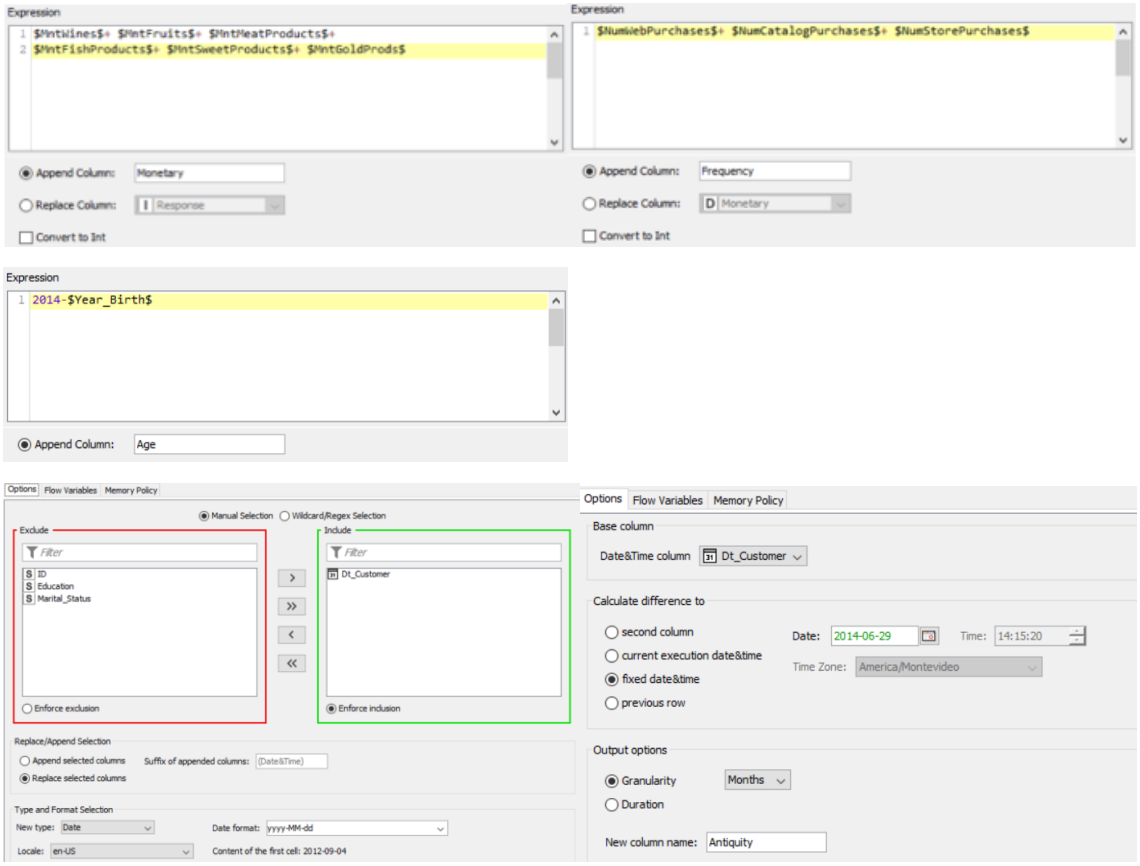
Entrenamiento	
Variables	
Variables no ausentes	No
Corte ausente	50.0
Variables de clase	
Método de imputación predeterm	Distribución
Método predeterminado de la var	Ninguno
Normalizar valores	Sí
Variables de intervalo	
Método de imputación predeterm	Distribución
Método predeterminado de la var	Ninguno
Valor constante predeterminado	
Valor alfanumérico predetermina	
Valor numérico predeterminado	.
Opciones de método	
Semilla aleatoria	12345
Parámetros de ajuste	
Imputación de árbol	
Puntuación	
Ocultar variables originales	Sí
Variables de indicador	
Tipo	Ninguno
Fuente	Variables imputadas
Rol	Rechazado

7.2. Capturas configuración nodos utilizados en KNIME

La herramienta KNIME se utilizó principalmente para la preparación de las variables.

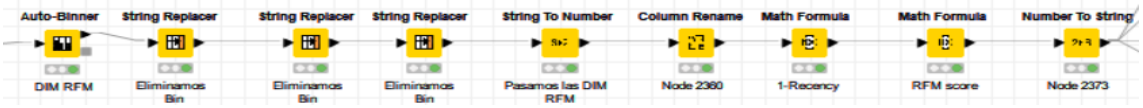


En primer lugar, se crearon las variables Age, Antiquity, Monetary, Frequency con el nodo Math formulas, la forma que se crearon fue la siguiente:

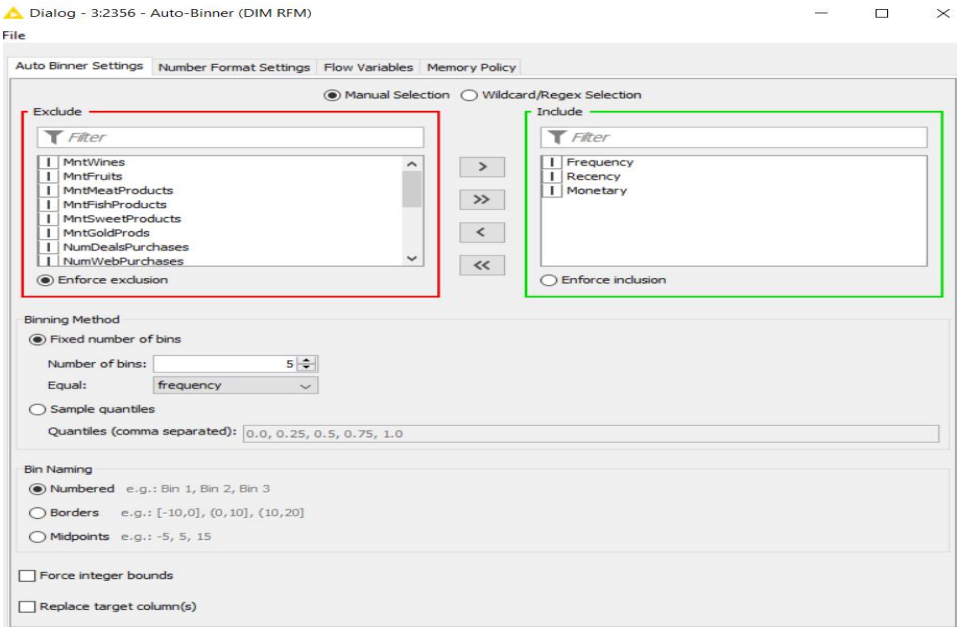


Una vez creadas las variables Age, Antiquity y Response, las variables Year_Birth, Dt_Customer, AcceptedCmp fueron eliminadas con un nodo column filter.

Para el RFM Score:

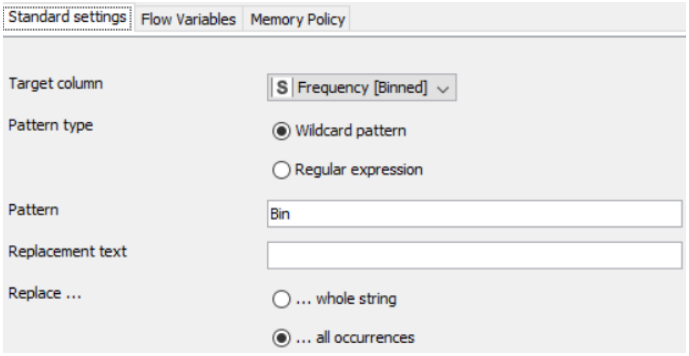


Primero se crearon los intervalos de las tres dimensiones del modelo RFM. Se hizo con el nodo Auto-Binner y el mismo fue configurado de la siguiente manera:



El Equal = Frequency es fundamental para que todos los intervalos tengan la misma cantidad de observaciones.

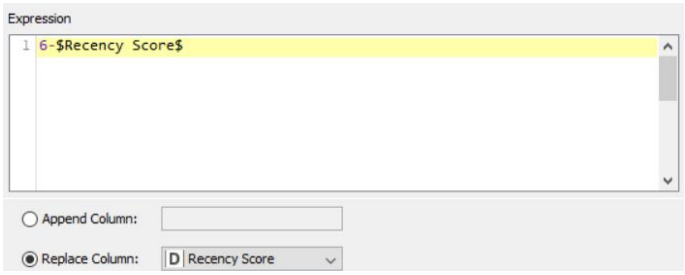
El auto binner por defecto a cada intervalo le pone bin 1, bin 2 … bin 5. Como los bin son creados de forma ordenada y ascendente, (en el bin 1 se encuentran los valores más bajos y en el 5 los más altos). Quitando la palabra bin de las observaciones, ya obtenía el Frequency score, Monetary Score y Recency score:



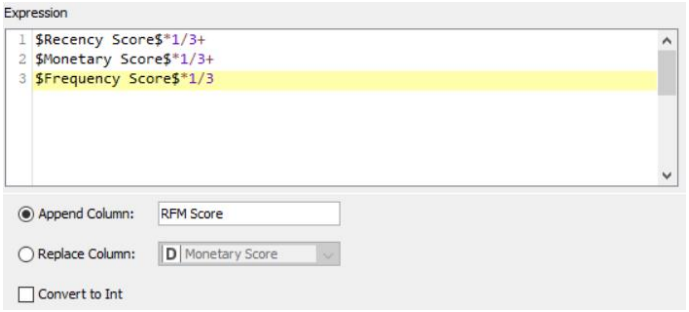
Se hizo lo mismo para las tres dimensiones.

Una vez que teníamos los scores para las tres dimensiones, se convirtieron las variables de categóricas a numéricas.

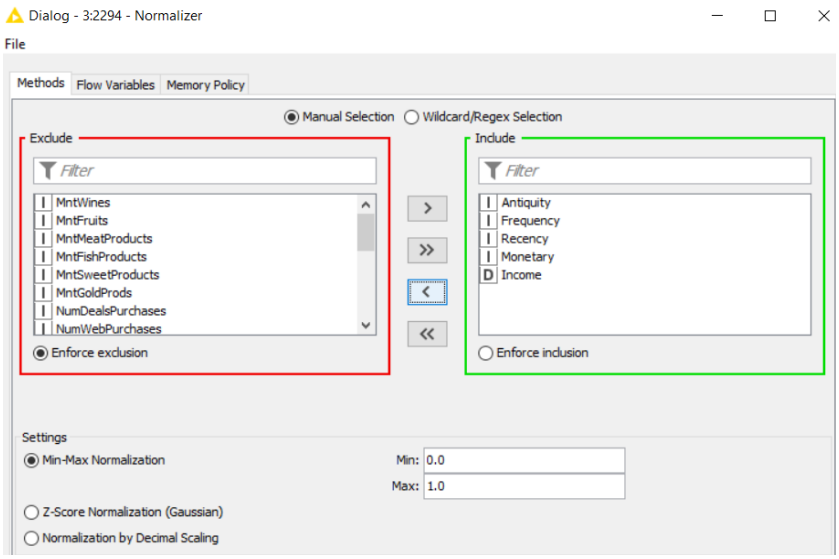
En el caso de Recency, que 1 pasó a ser las observaciones con valores más pequeños, lo que se hizo fue invertir la variable (nosotros queremos que la persona con Recency Score de 1 sea el que tiene recencia más alta). Esto se hizo con un nodo Math formula:



Una vez que teníamos los tres scores correctamente, se creó el RFM score, dándole la misma importancia a las tres dimensiones:

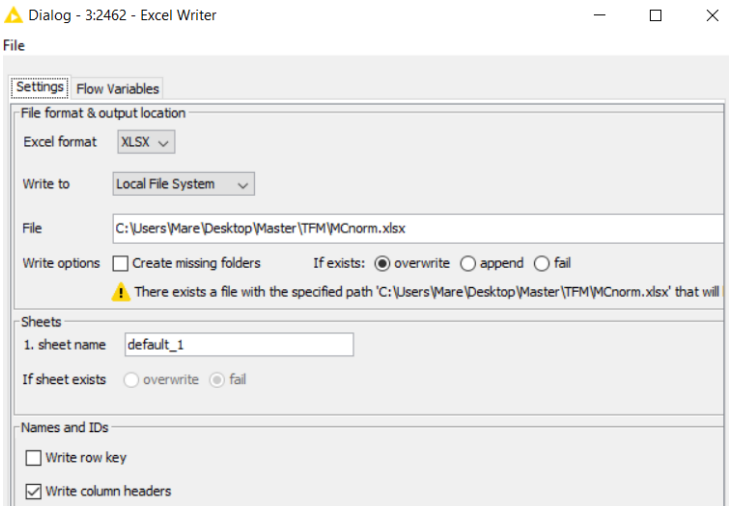


El objetivo tres y cuatro del proyecto era realizar segmentaciones de clientes. Para ello debíamos tener las variables normalizadas. Para la segmentación en base a las tres dimensiones de del modelo RFM, se podrían haber utilizado los Scores individuales. De todas maneras, yo preferí trabajar con las variables originales normalizadas.

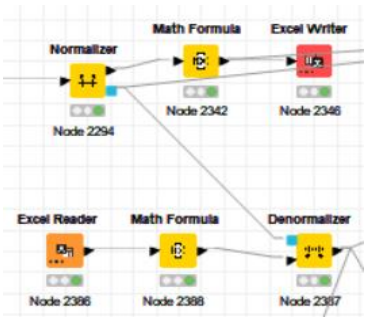


Se puede ver que se utilizó tipo de normalización min max. Se normalizaron únicamente las variables que se iban a utilizar: Antiquity, Frequency, Recency, Monetary y Income.

Una vez que tenía las variables normalizadas, la base de datos fue guardada con un nodo Excel Writer con el nombre de MCnorm para poder trabajar en R con los datos listos.



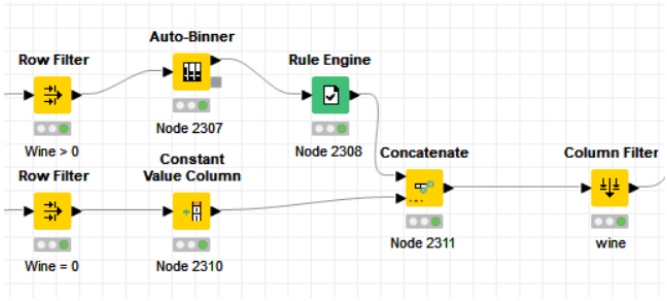
Con los clústeres ya formados, se volvieron a llevar los datos a KNIME, para estudiarlos con las variables desnormalizadas. Se hicieron gráficos de tartas y se sacaron conclusiones:



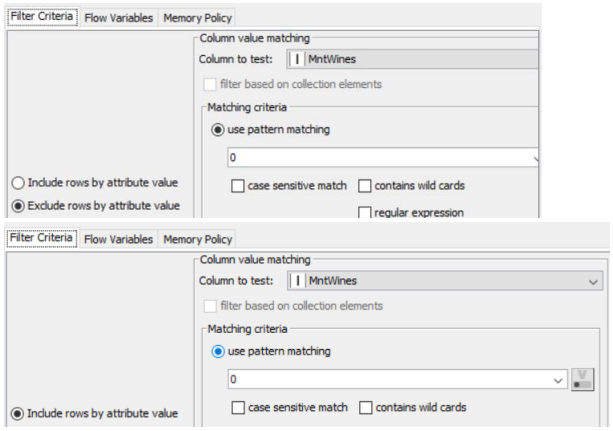
Se conectó el nodo denormalizer al nodo Normalizer que se había utilizado para normalizar las variables en primer lugar. El nodo Math formula, previo al nodo denormalizer es para revertir el cambio que se le había hecho a la variable Recency (1-recency).

Para el objetivo 5, en donde se buscaba identificar el perfil de los clientes con la utilización del algoritmo a priori, se necesitaba que todas las variables fueran "nominales". Por esta razón, las variables numéricas, había que transformarlas en categóricas mediante la creación de intervalos.

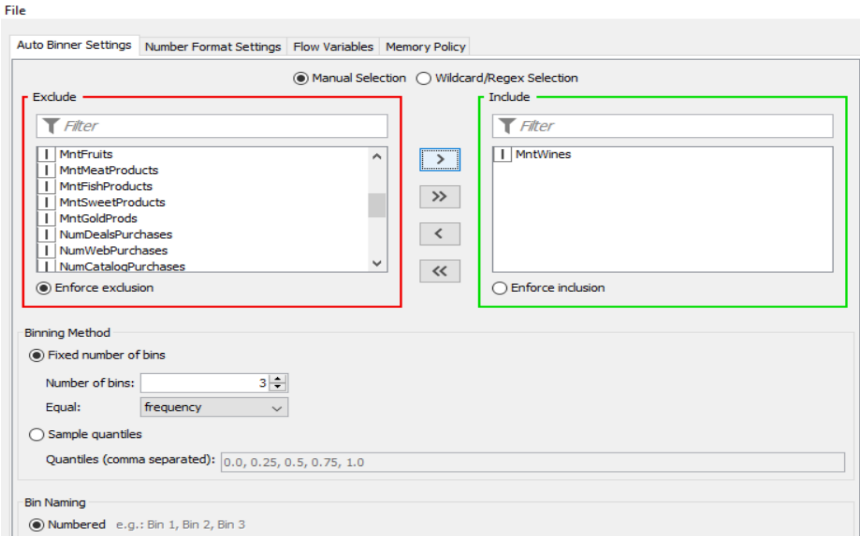
Para la creación de los intervalos de las variables "Mnt…" que indican el consumo de cada uno de los productos, los intervalos fueron creados de la siguiente manera (se va a poner el ejemplo de los vinos, pero todas las variables de cada uno de los 6 productos fueron preparadas de la misma manera):



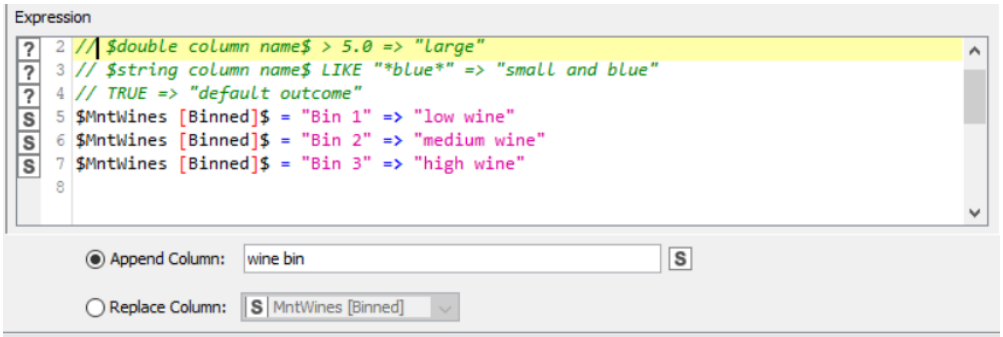
En primer lugar, se utilizaron dos nodos Row filter, uno de ellos para filtrar y quedarnos únicamente las observaciones con valores distintos de 0 y otro para filtrar y quedarnos únicamente con los clientes que no consumieron vino.



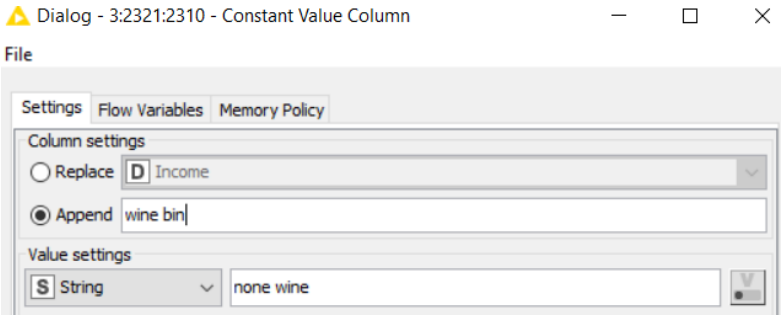
A las observaciones distintas de 0, se les aplico un nodo Auto binner y se dividió el total de las observaciones distintas de 0 en tres intervalos con la misma cantidad de observaciones.



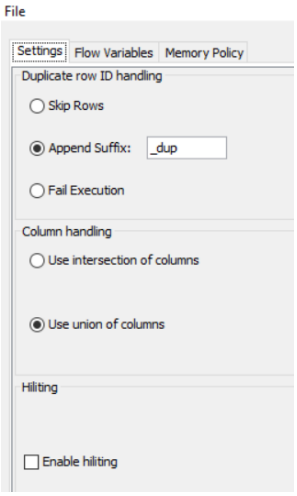
Una vez que obtuve los tres intervalos, con el nodo Rule Engine se renombraron las observaciones. Todas las observaciones que tomaran el valor “Bin 1”, pasaron a tomar el valor “low wine”, y así. En lugar de reemplazar la columna que se había creado, se creó una nueva llamada “wine bin”:



Por otra parte, a las observaciones que habían sido filtradas por no haber consumido vino, se les agregó una nueva columna llamada “wine bin” que tomara el valor constante “none wine”



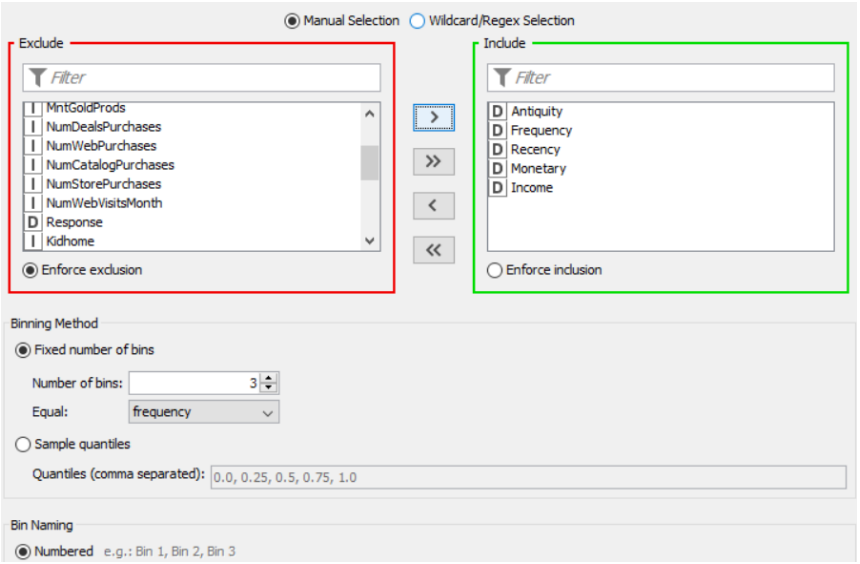
Una vez que todas las observaciones tenían un valor para “wine bin” se concatenaron las bases de datos con el nodo “concatenate”



Luego se excluyeron de la base de datos las columnas que no iban a ser utilizadas con la utilización del nodo columna filter.

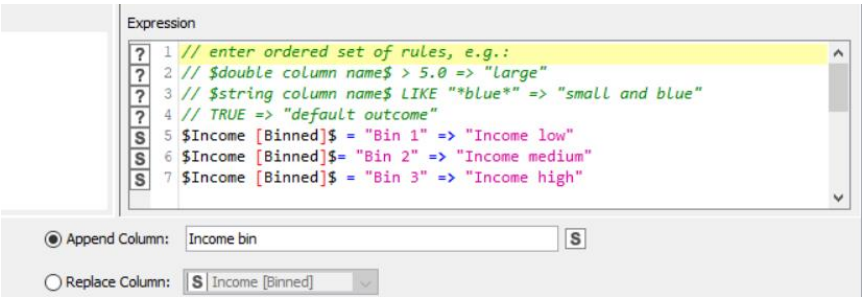
Este proceso se repitió con todos los productos de la base de datos.

Luego se crearon los intervalos de las variables Antiquity, Frequency, Recency, Monetary e Income:



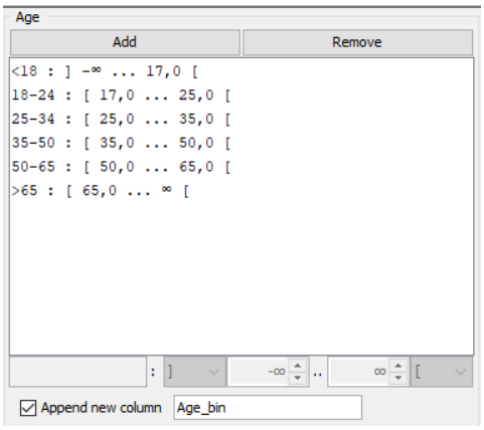
Todas estas variables se dividieron en tres intervalos con la misma frecuencia.

Una vez que fueron creados, también con la utilización del nodo rule Engine, los valores fueron renombrados:

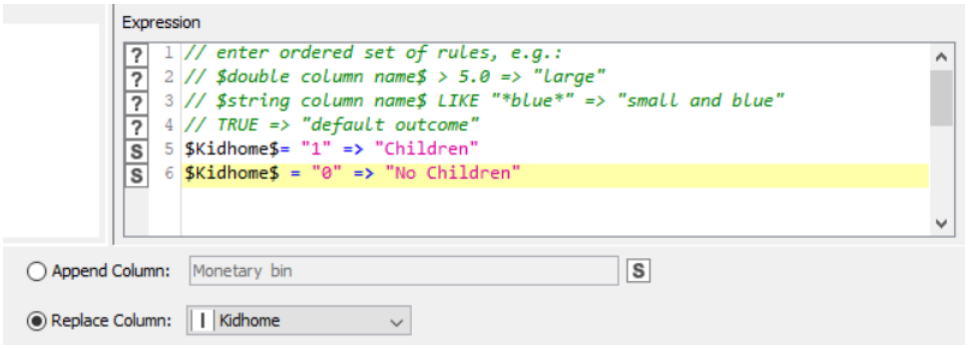


Se hizo lo mismo en todas las variables.

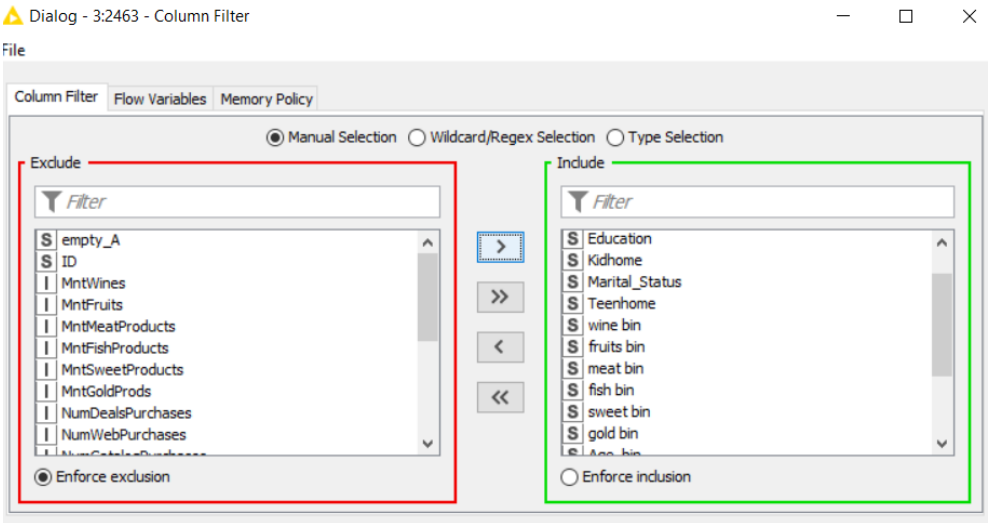
Para crear los intervalos de la variable Age, se utilizó un numeric binner, y los intervalos fueron creados de forma manual, y no proporcionalmente:



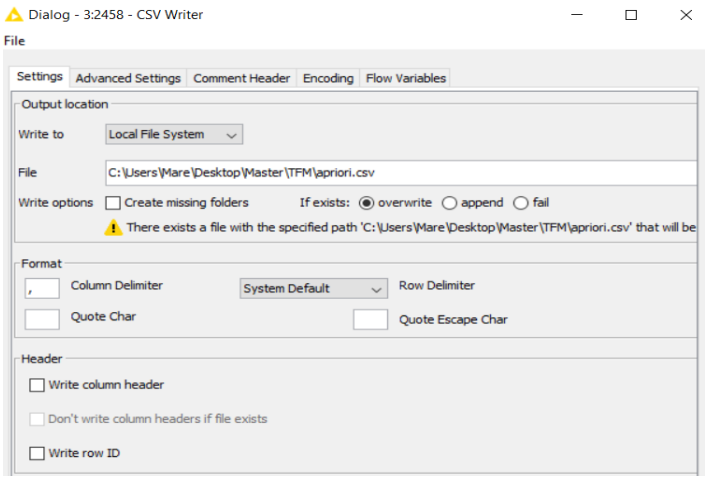
Los valores de las variables Teenhome y Kidhome, fueron remplazados por nuevas expresiones:



Para poder aplicar el algoritmo a priori en R, antes de guardar los datos para poder importarlos en R, se aplicó un column filter para quedarme únicamente con las variables categóricas que iba a querer usar en el algoritmo a priori:



Los datos fueron guardados en un archivo cvs con el nombre apriori:



Se guardaron sin el nombre de las columnas ya que para utilizar el algoritmo a priori no son necesarios.

7.3. Código R

7.3.1. Clustering

```
install.packages("factoextra")
install.packages("clustertend")
install.packages("NbClust")
install.packages("xlsx")

library(factoextra)
library(clustertend)
library(NbClust)
library(readxl)
library(xlsx)

df <- read_excel("C:/Users/Mare/Desktop/Master/TFM/MCnorm.xlsx")
view(df)

# Modelo RFM

dput(names(df))
RFM <- df[ ,c("Recency", "Frequency", "Monetary")]
mon <- df[ ,c("MntWines", "MntFruits", "MntMeatProducts",
              "MntFishProducts", "MntSweetProducts", "MntGoldProds")]

# Se visualizan boxplot para ver la distribucion de cada dimension del modelo RFM y para ver la distirbucion de la cantidad
#consumida de cada producto
par(cex.axis=0.95,las=1)
boxplot(mon, col = rainbow(ncol(mon)))
boxplot(RFM, col = rainbow(ncol(mon)))

# Estadistico Hopkins:
res <- get_clust_tendency(RFM, n = nrow(RFM)-1, graph = TRUE)
res

# Se realiza PCA y se gráfica para observar los datos

fviz_pca_ind(prcomp(RFM), title = "PCA - RFM",
             geom = "point", ggtheme = theme_classic())

# En caso que se aplicaran 4 clústeres con el algoritmo k-media se vería asi:

km.res1 <- kmeans(RFM, 4, iter.max = 10, nstart = 25)
fviz_cluster(list(data = RFM, cluster = km.res1$cluster),
             ellipse.type = "norm", geom = "point", stand = FALSE,
             palette = "jco", ggtheme = theme_classic())


# Numero optimo de clusters

# Elbow
fviz_nbclust(RFM, kmeans, method = "wss") +
  geom_vline(xintercept = 4, linetype = 2)
 labs(subtitle = "Elbow method")

# Silhouette method
fviz_nbclust(RFM, kmeans, method = "silhouette")+
  labs(subtitle = "Silhouette method")

#R ofrece un paquete que busca el número optimo de clusteres por 30 metodos distintos. Luego grafica cuantas veces fue
#elegido ese numero de clusteres.


nb <- NbClust(RFM, distance = "euclidean", min.nc = 2,
              max.nc = 10, method = "kmeans")
fviz_nbclust(nb)

# 4 cluteres es el mas elegido
```

```r
# k=4
km.RFM4 <- eclust(RFM, "kmeans", k = 4, nstart = 25, graph = FALSE)
fviz_cluster(km.RFM4, data = RFM,
             ellipse.type = "euclid",
             star.plot = TRUE,
             repel = TRUE,
             ggtheme = theme_minimal()
)

fviz_cluster(km.RFM4, geom = "point", ellipse.type = "norm",
             palette = "jco", ggtheme = theme_minimal())

fviz_silhouette(km.RFM4, palette = "jco",
                ggtheme = theme_classic())


dfRFM <- cbind(RFM, "cluster" = km.RFM4$cluster)
print(dfRFM)
par(cex.axis=1)
boxplot(data=dfRFM,Recency~cluster,main="Rececncy", col = rainbow(ncol(dfRFM)))
boxplot(data=dfRFM,Monetary~cluster,main="Monetary", col = rainbow(ncol(dfRFM)))
boxplot(data=dfRFM,Frequency~cluster,main="Frequency", col = rainbow(ncol(dfRFM)))


# K=5
km.RFM5 <- eclust(RFM, "kmeans", k = 5, nstart = 25, graph = FALSE)

fviz_cluster(km.RFM5, data = RFM,

             ellipse.type = "euclid",
             star.plot = TRUE,
             repel = TRUE,
             ggtheme = theme_minimal()
)

fviz_cluster(km.RFM5, geom = "point", ellipse.type = "norm",
             palette = "jco", ggtheme = theme_minimal())

fviz_silhouette(km.RFM5, palette = "jco",
                ggtheme = theme_classic())

km.RFM5$centers


#K=6
km.RFM6 <- eclust(RFM, "kmeans", k = 6, nstart = 25, graph = FALSE)
fviz_cluster(km.RFM6, data = RFM,
             ellipse.type = "euclid",
             star.plot = TRUE,
             repel = TRUE,
             ggtheme = theme_minimal()
)


fviz_cluster(km.RFM6, geom = "point", ellipse.type = "norm",
             palette = "jco", ggtheme = theme_minimal())


fviz_silhouette(km.RFM6, palette = "jco",
                ggtheme = theme_classic())

km.RFM6$centers


km.RFM7 <- eclust(RFM, "kmeans", k = 7, nstart = 25, graph = FALSE)

fviz_cluster(km.RFM7, data = RFM,

             ellipse.type = "euclid",
             star.plot = TRUE,
             repel = TRUE,
             ggtheme = theme_minimal()
)

fviz_cluster(km.RFM7, geom = "point", ellipse.type = "norm",
             palette = "jco", ggtheme = theme_minimal())

fviz_silhouette(km.RFM7, palette = "jco",
                ggtheme = theme_classic())

km.RFM7$centers

dfRFM7 <- cbind(RFM, "cluster" = km.RFM7$cluster)
print(dfRFM7)
par(cex.axis=1)
boxplot(data=dfRFM7,Recency~cluster,main="Rececncy", col = rainbow(ncol(dfRFM7)))
boxplot(data=dfRFM7,Frequency~cluster,main="Frequency", col = rainbow(ncol(dfRFM7)))
boxplot(data=dfRFM7,Monetary~cluster,main="Monetary", col = rainbow(ncol(dfRFM7)))

dftot <- cbind(df, "cluster4" = km.RFM4$cluster, "cluster7" = km.RFM7$cluster )
write.xlsx(dftot, "dftot.xlsx")
```

```
############ Segunda segmetnación ##############

dput(names(df))
obj2 <- df[ ,c("Income", "Antiquity", "Monetary")]

# Se visualizan boxplot para ver la distribucion de cada dimension del modelo obj2 y para ver la distirbucion
#de la cantidad consumida de cada producto
par(cex.axis=0.95,las=1)
boxplot(obj2, col = rainbow(ncol(mon)))

# Estadistico Hopkins:
res2 <- get_clust_tendency(obj2, n = nrow(obj2)-1, graph = TRUE)
res2

# Se realiza PCA y se gráfica para observar los datos

fviz_pca_ind(prcomp(obj2), title = "PCA - obj2",
             geom = "point", ggtheme = theme_classic())

# En caso que se aplicaran 4 clústeres con el algoritmo k-media se vería asi:

km.res2 <- kmeans(obj2, 4, iter.max = 10, nstart = 25)
fviz_cluster(list(data = obj2, cluster = km.res2$cluster),
             ellipse.type = "norm", geom = "point", stand = FALSE,
             palette = "jco", ggtheme = theme_classic())
km.res2$centers


# Numero optimo de clusters

# Elbow
fviz_nbclust(obj2, kmeans, method = "wss") +
  geom_vline(xintercept = 4, linetype = 2)
labs(subtitle = "Elbow method")

# Silhouette method
fviz_nbclust(obj2, kmeans, method = "silhouette")+
  labs(subtitle = "Silhouette method")

#R ofrece un paquete que busca el número optimo de clusteres por 30 metodos distintos. Luego grafica cuantas
#veces fue elegido ese numero de clusteres.

nb2 <- NbClust(obj2, distance = "euclidean", min.nc = 2,
               max.nc = 10, method = "kmeans")
fviz_nbclust(nb2)


# k=4

km.obj24 <- eclust(obj2, "kmeans", k = 4, nstart = 25, graph = FALSE)
fviz_cluster(km.obj24, data = obj2,
             ellipse.type = "euclid",
             star.plot = TRUE,
             repel = TRUE,
             ggtheme = theme_minimal()
)

fviz_cluster(km.obj24, geom = "point", ellipse.type = "norm",
             palette = "jco", ggtheme = theme_minimal())

fviz_silhouette(km.obj24, palette = "jco",
                ggtheme = theme_classic())

km.obj24$centers

dfobj2 <- cbind(obj2, "cluster" = km.obj24$cluster)
print(dfobj2)
par(cex.axis=1)
boxplot(data=dfobj2,Income~cluster,main="Income", col = rainbow(ncol(dfobj2)))
boxplot(data=dfobj2,Monetary~cluster,main="Monetary", col = rainbow(ncol(dfobj2)))
boxplot(data=dfobj2,Antiquity~cluster,main="Antiquity", col = rainbow(ncol(dfobj2)))

dftot2 <- cbind(dftot, "cluster_obj4" = km.obj24$cluster)

write.xlsx(dftot2, "dftot2.xlsx")


# k=5

km.obj25 <- eclust(obj2, "kmeans", k = 5, nstart = 25, graph = FALSE)
fviz_cluster(km.obj25, data = obj2,
             ellipse.type = "euclid", # Concentration ellipse
             star.plot = TRUE, # Add segments from centroids to items
             repel = TRUE, # Avoid label overplotting (slow)
             ggtheme = theme_minimal()
)

fviz_cluster(km.obj25, geom = "point", ellipse.type = "norm",
             palette = "jco", ggtheme = theme_minimal())

fviz_silhouette(km.obj25, palette = "jco",
                ggtheme = theme_classic())

km.obj25$centers
```

7.3.2. Algoritmo a priori

```
library(readxl)
apriori <- read_excel("C:/Users/Mare/Desktop/Master/TFM/apriori2.xlsx")

install.packages("arules")
library("arules")

# Para poder realizar el analisis de asociaciones es necesario crear un conjunto de datos transaccional:

caracteristicas <- read.transactions("apriori.csv", format="basket", sep= ",", header= FALSE)
inspect(caracteristicas)

# Se va a inspeccionar el archivo caracteristicas.

itemFrequencyPlot(caracteristicas, topN=10, type= 'absolute')
tamanyos <- size(caracteristicas)
summary(tamanyos)

library(dplyr)
frecuencia_items <- itemFrequency(x = caracteristicas, type = "relative")
frecuencia_items %>% sort(decreasing = TRUE) %>% head(20)

frecuencia_items <- itemFrequency(x = caracteristicas, type = "absolute")
frecuencia_items %>% sort(decreasing = TRUE) %>% head(5)

dim(caracteristicas)

# Se van a buscar los "itemsets" frecuentes. En este caso lo que se hace es buscar caracteristicas que se den frecuentemente juntas.
# Por ejemplo, persona entre 35 y 50 años con ingreso alto y con niños en el hogar
# Se va a pedir que los itemsets encontrados al menos tengan dos items

itemsets <- apriori(data = caracteristicas,
                    parameter = list(support = 0.1,
                                     minlen = 2,
                                     maxlen = 20,
                                     target = "frequent itemset"))
summary(itemsets)

# Si vemos los 20 itemnsets mas frecuentes:

top_20_itemsets <- sort(itemsets, by = "support", decreasing = TRUE)[1:20]
inspect(top_20_itemsets)


# Ahora se va a pasar a observar reglas de asociación

reglas <- apriori(data = caracteristicas,
                  parameter = list(support = 0.1,
                                   confidence = 0.60,
                                   # Se especifica que se creen reglas
                                   target = "rules", minlen=2))
summary(reglas)

# Se van a ordenar las reglas en funcion de la confianza para que nos muestre las reglas con mayor confianza primero:

reglas <- sort(reglas, by="confidence", decreasing = TRUE)

str(reglas)
inspect(reglas)

# Si observamos las reglas maximales:
# (se consideran reglas maximales cuando no hay otro itemset que sea superset. Es decir es el itemset con mas itmes posible)
reglas_maximales <- reglas[is.maximal(reglas)]
reglas_maximales

inspect(reglas_maximales)

# Si se quisiera podar para quitar las reglas redundantes
# Una regla redundante si cubre la misma información, o información menos general, que la información que cubre otra regla de la misma
# utilidad y relevancia (misma o mas confianza)
reglas_redundantes <- is.redundant(reglas,  measure = "confidence")
reglas_redundantes
inspect(reglas[is.redundant(reglas)])

which(is.redundant(reglas))
reglas_podado <- reglas[!reglas_redundantes]
reglas_podado <- sort(reglas_podado, by="confidence", decreasing = TRUE)

inspect(reglas_podado)


# si ahora quiero pasar a mirar los itemnsets que contengan la caracteristica "high wine"

itemsets_wine <- arules::subset(itemsets,
                                subset = items %in% "high wine")
itemsets_wine
inspect(itemsets_wine[1:10])

# Ahora se van a buscar aquellas reglas en donde el consecuente sea high wine
# Con esto se van a intentar encontrar las caracteristicas que llevan a un consumo alto de vino

wines <- apriori(data = caracteristicas,
                             parameter = list(support = 0.1,
                                              confidence = 0.70,
                                              # Se especifica que se creen reglas
                                              target = "rules"),
                             appearance = list(rhs = "high wine"))

wines <- sort(wines, by="confidence", decreasing = TRUE)

inspect(wines[1:30])

# Si observamos las reglas maximales:
# (se consideran reglas maximales cuando no hay otro itemset que sea superset. Es decir es el itemset con mas itmes posible)
reglas_maximales <- wines[is.maximal(wines)]
reglas_maximales

inspect(reglas_maximales[1:10])
```

```r
# Si se quisiera podar para quitar las reglas redundantes
# Una regla redundante si cubre la misma información, o información menos general, que la información que cubre otra regla de la misma
#utilidad y relevancia (misma o mas confianza)
reglas_redundantes <- is.redundant(wines,  measure = "confidence")
reglas_redundantes
inspect(wines[is.redundant(wines)])

which(is.redundant(wines))
wines_podado <- wines[!reglas_redundantes]
wines_podado <- sort(wine_podado, by="confidence", decreasing = TRUE)

inspect(wines_podado)

# Si ahora queremos filtrar que en el consecuente se de el mayor consumo de vino, pero el antecedente que solo dependa de las
# caracteristicas personales y no de los habitos de compra,(considero que estan muy relacionadas, si tiene alto monetary y frequency,
#va a ser high wine)

winesper <- apriori(data=caracteristicas, parameter = list(support = 0.1,
                                                           confidence = 0.60,
                                                           # Se especifica que se creen reglas
                                                           target = "rules",  minlen=2),
                    appearance = list(rhs = "high wine", lhs = c("No Children", "Children", "No Teens", "Teens",
                                                                 "Income high", "Income medium", "Income low", "PhD", "Master"
                                                               , "Graduation", "2n Cycle", "Single", "In couple", "18-24", "25-34",
                                                              "35-50", "50-65", ">65")))
winesper <- sort(winesper, by="confidence", decreasing = TRUE)
inspect(winesper)

#### REPITO LO MISMO PARA TODOS LOS PRODUCTOS #####
```